International Journal on Advances in Systems and Measurements

2025 vol. 18 nr. 3&4

The *International Journal on Advances in Systems and Measurements* is published by IARIA.
ISSN: 1942-261x
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Systems and Measurements
Volume 18, Number 3&4, 2025

International Journal on Advances in Systems and Measurements
Volume 18, Numbers 3&4, 2025

CONTENTS

The Generality-Accuracy Trade-off in Neural State Estimation

Aleksandr Berezin, Eric MSP Veith and Stephan Balduin
R&D Division Energy
OFFIS – Institute for Information Technology
Oldenburg, Germany
e-mail: {aleksandr.berezin | eric.veith | stephan.balduin}@offis.de

Thomas Oberließen and Sebastian Peter
Institute of Energy Systems, Energy Efficiency and Energy Economics
Technical University of Dortmund
Dortmund, Germany
e-mail: {thomas.oberliessen | sebastian.peter}@tu-dortmund.de

***Abstract*—This paper addresses the challenge of neural state estimation in power distribution systems. We identified a research gap in the current state of the art which lies in the inability of models to adapt to changes in the power grid, such as loss of sensors and branch switching, in a zero-shot fashion. We designed benchmarks to evaluate the robustness of models to different changes in grid topology and used them to test models with different architectures. The observed results strongly suggest the existence of a trade-off between accuracy and robustness.**

Keywords-neural state estimation; zero-shot learning; transfer learning; graph neural networks.

I. Introduction

This work extends the results of our conference paper [1], which considered the problem of zero-shot Neural State Estimation (NSE) with a focus on the relationship between model complexity and performance. This extension provides better benchmarks and evaluations, focusing on the apparent trade-off between accuracy and generality of NSE models. It also expands the set of tested models to include non-graph-based architectures and a non-parametric baseline.

We begin by reviewing relevant prior work in Section II, followed by a formal statement of our research question in Section III. Section IV details the methodology, model selection, and experimental setup. The results of the experiments are presented in Section V. Finally, Section VI summarizes our findings and suggests directions for future work.

II. State of the art

Power System State Estimation (PSSE) is the task of inferring the "state" of an electrical power grid from real-time data collected by various sensors distributed throughout the system. The "state" in this context generally refers to the voltage magnitudes and phase angles at each bus in the grid.

For many years, PSSE was mainly performed for transmission grids using simplifying assumptions such as near-DC power flow and computational methods with poor scalability [2]. This is enabled by balanced operation with a relatively simple, predominantly linear topology of transmission grids, given their scale and structure.

This approach cannot be extended to distribution grids that transport electricity from substations to end consumers. Their unbalanced nature, radial or weakly meshed topology, high R/X ratios, and above all, cost inefficiency to achieve sufficient sensor coverage complicate the state estimation process. Initially designed with transmission systems in mind, conventional methods often struggle to provide accurate state estimation in more complex, dynamic, and less predictable distribution systems [2].

However, with the proliferation of Distributed Energy Resources (DERs) and other complex consumers, grid operators are faced with the need to perform PSSE for distribution grids. Furthermore, §14a of the German Energy Industry Act effectively requires operators to develop observability in distribution grids in order to align consumption with production from renewable energy sources, which requires PSSE.

A. Conventional methods

The traditional and most widely used approach for PSSE is the Weighted Least Squares (WLS) method [3]. This algorithm minimizes the sum of the squared differences between the observed and estimated measurements, with each term being weighed inversely proportionally to the square of the measurement error standard deviation.

What limits the direct application of WLS in distribution systems is the minimum number of measurements required for the convergence of WLS. Assuming the grid contains n buses, it is then described by $2n$ variables, namely n voltage magnitude values and n voltage angles. A slack bus serves as the reference; its voltage angle is set to zero or a known constant, and therefore does not need to be estimated. The voltage angles of the other network buses are relative to the voltage angle of the connected slack bus. Therefore, the state estimation must find $2n-k$ variables, where k is the number of defined slack buses. The minimum amount of measurements m_{min} needed for the WLS method to work is therefore:

$m_{min} = 2n - k$

However, in order to perform well, the number of redundant measurements should be higher. A value of $m \approx 4n$ is often considered reasonable for practical purposes. This level of observability is unachievable in distribution grids due to

economic constraints and the sheer number of elements that must be monitored.

Another problem is that the WLS algorithm is computationally intensive. Assuming a dense system matrix, its time complexity is generally considered to be $\mathcal{O}(N^3)$, where N is the number of buses. This is due to the need for matrix inversions and solving linear equations. This complexity becomes a limitation for large-scale distribution grids with thousands of buses, leading to significant computational burden and time constraints, especially when real-time estimates are required. Additionally, WLS assumes that all error distributions are Gaussian, a condition that may not always hold in practice.

B. Feed-forward methods

The most promising path to overcome these limitations and provide observability in distribution grids is currently believed to be NSE: data-driven methods that utilize historical data in addition to real-time measurements. Artificial Neural Networks (ANNs) may be able to perform the calculation faster while being robust to insufficient measurements [4][5]. However, like all Machine Learning (ML) methods, the performance of ANNs is contingent on the quality and quantity of the available training data. Therefore, NSE approaches are usually valid only for the grid they have been trained on. Once the topology or characteristics of nodes change, the ANN needs to be retrained. This is known as the problem of Transfer Learning (TL).

When designing an ANN-based system for solving PSSE, it is tempting to start with Multi-Layer Perceptrons (MLPs). Not only are they the easiest ANNs to implement, but they can approximate any function guaranteed by the universal approximation theorem. In practice, it means that, given sufficient computing power and training data, such models can achieve an arbitrarily high level of accuracy. However, feed-forward models underutilize two important properties of power grids. The first property is that the grid can be represented as a graph. The second is that a power grid, like any physical system, is local, meaning that any interaction between two elements must propagate through the lines and buses between them. Together, these properties enable a drastic reduction in the number of possible interactions that a model needs to consider. The feed-forward models learn about this locality implicitly, inferring the grid topology from the data. However, this is a disadvantage when a topology change occurs, because the model's knowledge of the grid structure and the physical processes (graph signals) are entangled, leading to huge performance degradation when anything in the underlying system changes.

The same is true for other fully-connected feed-forward models that build upon the MLP, such as transformers; hence the title of this subsection.

C. Geometric methods

A sensible way to overcome this limitation is to use models that incorporate information about the graph topology into their calculations. Such models are known under an umbrella term Graph Neural Networks (GNNs), or, less commonly, "geometric models". They are specifically designed to separate the graph structure from the graph signals and only model the latter. This means that they also require the topology of the power grid as input. This can be a barrier if that topology is not known, as in this case a separate topology estimation system is required. However, this design drastically reduces the number of trainable parameters and allows a model trained on one graph to perform inference on another with little to no adjustments [1]. In the context of power systems, there are situations where the grid topology changes due to alterations in switch states or maintenance of elements, up to and including islanding, when parts of the grid become isolated from the greater system. Some of these events can occur suddenly and the control system must adapt to them in real time, which necessitates an ability to generalize provided by GNNs. However, the downside is that GNNs cannot generally approximate an arbitrary function, which in practice limits their maximum accuracy.

In this study, we will be using a subset of GNNs known as Message Passing Networks (MPNs). The most common examples of MPNs are Graph Convolutional Networks (GCNs), Graph Attention Networks (GATs), and Graph Isomorphism Networks (GINs). All of these models are built on the graph message passing operation, which has the locality property with respect to the graph geometry.

Expectedly, recent years have seen a high volume of publications that propose utilizing GNNs for NSE in various ways.

A recent study on GNNs for state estimation [6] came out of a collaboration between TenneT and Radboud University. It demonstrates a GCN-based topology control system to mitigate grid congestion. The model, trained on historical fault data, dynamically reconfigures the grid topology by opening or closing circuit breakers in response to overload warnings.

A similar project combining state estimation with active control is described in [7]. The researchers aimed to develop data analytics services for predicting localized grid congestion caused by excessive distributed renewable generation and eventually prevent it by issuing bids for purchasing energy flexibility on the market. To achieve this, they used a GNN model for both state estimation (to detect congestion) and for generating control signals (in the form of bids). The data used are live voltages and energy profiles from prosumers with PV systems, as well as the known grid topology. They note that GNNs are far more efficient than other tested models and are more capable of adapting to grid changes, while being slightly less accurate.

However, to our knowledge, none of these research projects specifically studied the problem of Zero-Shot Learning (ZSL) in PSSE. The contribution of this work is in setting up multiple evaluation scenarios for ZSL in NSE and using them to evaluate the performance of GNNs against other models.

D. Theoretical limitations of GNNs

While our paper investigates the capabilities and limitations of GNNs experimentally, several notable papers take a more rigorous route of building a theoretical understanding of them.

The capability that is critical to our use case is generalization across different topologies, which is analyzed by two papers.

The first study to successfully incorporate non-trivial graph similarities, architectural choices, and loss functions into generalization analysis is [8]. They analytically derive the bounds of generalization for GNNs as a function of these factors. The theoretical results are then verified on real datasets. The conclusion relevant to this study is that the generalization ability of GNNs leverages the correlation between graph structure and node labels.

A later study [9] pushes the theoretical analysis further by incorporating model complexity into the calculation of generalization bounds and shows that there is a trade-off between the generalization capability of a model and its complexity. However, increasing complexity does not necessarily degrade generalization if it aligns with the task at hand. Quantifying this alignment is another major contribution of the study, as it gives new tools for choosing better GNN configurations for a given problem.

However, model complexity also affects performance of GNNs in scenarios without topology changes. This brings us to another important capability of GNNs: scaling, i.e., the possibility of increasing the accuracy of the model by adding more layers, as is usually done in feed-forward models.

Unfortunately, this mode of scaling in GNNs is limited by the oversmoothing phenomenon [10]. It is a consequence of GNN layers acting as low-pass filters, which effectively averages the output values over multiple iterations. Eventually, the model converges to an output where the values at all nodes of the graph are identical. Therefore, GNN models have a finite optimal depth that can differ between graphs and model architectures and, therefore, is usually found empirically.

A deep analysis of the oversmoothing phenomenon is presented in [11]. They find that the onset of oversmoothing is related to the graph diameter, which is usually small for real-life graphs. After the number of layers surpasses the diameter, for each node, there will be no nodes that have not been encountered before in message passing and hence the node representations will tend to homogenize. In contrast to most other graph datasets, power grids are characterized by long linear branches and therefore large diameters, which should make oversmoothing less likely to occur. However, this structure presents challenges of its own for GNNs.

III. Research question

When discussing the ability of a model to generalize to different grid topologies, it is important to differentiate between *homogeneous* and *heterogeneous* modes of TL. In general, the homogeneous TL mode means that the source and target data are in the same feature space, while in the heterogeneous TL mode they are represented in different feature spaces.

In the context of power grids, this is the difference between two use cases. In the homogeneous case, the power grid remains the same, but some connections between its nodes appear or disappear due to changes in switch states or elements going in and out of service. In the heterogeneous case, the model trained on one grid is used to make predictions about a completely different grid [12].

This distinction becomes very important in real-life deployments. Integrating a new model into the control system of a real grid naturally takes time, and training the model on that specific grid could be incorporated into this process without noticeably slowing it down. On the other hand, changes in grid topology due to switching can happen suddenly and unpredictably, and the model must adapt to them in real time.

There is also another way in which the data distribution can shift in the context of PSSE: the observable subset of buses can change, which changes the amount of input data points available to the model. This can also be considered a form of homogeneous TL.

A subset of TL is Zero-Shot Learning (ZSL). This scenario excludes the possibility of fine-tuning the model on the new distribution and evaluates its performance directly after the transfer. In this project, we specifically focus on ZSL because it is more representative of real-life situations where a model must make predictions immediately after a topology change without access to any training data for fine-tuning. In other words, the model should be *robust* to distributional shifts.

Of course, in practice, a model can be fine-tuned to provide the best performance for the new topology. Still, until this process is complete, the previous version of the model has to substitute for it and provide sufficiently good estimates, even if they are of lower quality.

The research question for this paper is what existing models in application to the PSSE problem are robust to changes in the data distribution, specifically:

A To the reduction of the subset of observable buses;
B To grid topology changes resulting from changing switch states;
C To transfer to a completely different power grid.

IV. Methodology

Before we proceed to describe the models we investigate, please note that our model implementations may not be ideal and therefore may not provide the most accurate results in absolute terms. This study should not be taken as an attempt to rank different models and determine the best performing ones, but rather to observe the impact of graph topology changes on the models' performance. This metric should theoretically be more robust to imperfections in model implementations, since it reflects the more fundamental structural properties of the models in question.

A. GNN models

We are comparing four GNN models using the implementations provided by the PyTorch Geometric framework [13]:

1) Graph Convolutional Network (GCN) as proposed in [14]
2) Graph Attention Network (GAT) as proposed in [15]
3) Graph Isomorphism Network (GIN) as proposed in [16]
4) Graph Sample and Aggregate (GraphSAGE) as proposed in [17]

Each model is tested with a variable number of layers ranging from 1 to 10 as a hyperparameter. This is needed to empirically determine the optimal depth of a GNN where it is sufficiently expressive but not yet affected by oversmoothing. Later in Section V the models will be labeled by a concatenation of the architecture name with the number of layers, i.e., "GAT3" is a GAT model with 3 layers.

The models are trained to predict two features: the real and imaginary parts of the complex voltage for every node. The number of features in the hidden layers is the same. We use the Huber loss function [18], a dropout probability of 0.5 and the GraphNorm normalization method from [19]. The optimizer is Adam with a learning rate of 0.001.

B. MLP models

Although the most recent research on NSE focuses primarily on GNNs, the classic MLP architecture is also considered for this role, and experiments with them provide a valuable perspective.

One of the most cited models of this type is Physics-Aware Neural Network (PAWNN), proposed in [20]. The idea of it is to use the classic perceptron as a building block but prune its synapses according to the graph adjacency matrix, i.e., the grid topology. These perceptrons are stacked in a variable number of layers, equal to the maximum diameter of a vertex-cut partition of the original graph.

There also exists an improved derivation of this model, proposed in [21]. The improvement is based on the observation that designing the ANN architecture based on the adjacency matrix, as in the original PAWNN, may lead to unnecessary connections between layers. The Pruned Physics-Aware Neural Network (P2N2) cuts out those unnecessary connections and uses separate weight matrices for the individual parts of the ANN, depending on the grid topology.

Another approach is the Prox-Linear Network (PLN) model proposed in [22], which is based on a prox-linear solver for state estimation using the Least Absolute Value (LAV) method. The main idea is to split the nonlinear state estimation problem into several blocks that are proximally linear. The PLN is built by unfolding these blocks. In practice, this structure reduces to a MLP.

For this project, we reproduced the P2N2 and PLN models from their descriptions in the corresponding papers. This means that the implementation may not be entirely faithful to what the authors intended, but this is an unavoidable limitation caused by the lack of reference implementations.

C. Baseline

Choosing a baseline method for NSE is made difficult by the absence of a single commonly accepted method that works under the condition of partial observability (which excludes WLS). The solution we chose is the non-parametric feature propagation algorithm from [23], which interpolates missing node-level features by solving a heat equation with known features as boundary conditions. This results in a smooth interpolation of features between known nodes. Of course, this algorithm is not designed for PSSE and is not expected to perform well, but it gives a deterministic solution that is easy to grasp intuitively, which makes it a suitable baseline.

D. Graph representations of power systems

A successful application of GNN models naturally depends on how well the underlying data can be represented in the graph format. The most obvious representation, and the one used in this paper, is known as the bus-branch model. In it, buses are represented as nodes of the graph, while lines and transformers (branches) are its edges, with branch admittances as edge weights. Voltages normalized to local reference values are the node features.

Admittances are chosen as edge weights because the graph Laplacian operator assumes higher edge weights to mean a higher correlation between nodes. This operator is, in turn, used in both the GNN models and the feature propagation algorithm. It should be noted that the models in question support neither complex-valued weights nor multidimensional weights, so we have to use the magnitude of the true complex impedance.

However, using admittance instead of impedance as edge weights becomes a problem for representing closed switches, which have zero impedance and, therefore, infinite admittance. This problem is solved by fusing buses connected by closed bus-to-bus switches into one bus. This is complicated because multiple closed switches are often connected to the same bus, so a naive approach of fusing adjacent buses in random order does not work. Instead, we use an iterative algorithm inspired by [24]. Firstly, we build an auxiliary graph of just the closed bus-to-bus switches with buses as nodes and switches as edges. In this graph, nodes with a degree of one can be safely removed (fused with their adjacent buses). This, in turn, will lower the degree of the adjacent node. Eventually, every node will reach a degree of one and can be fused until every connected component of the auxiliary graph is fused into a single node.

However, it should be noted that the bus-branch model is not the only power grid representation that exists in the literature. For example, in [25], the authors used a more granular representation: they model each grid element as a separate node, with the addition of extremity nodes for connecting elements such as lines or transformers. The main limitation of this approach and the reason we choose not to use it is that it requires training data to include voltage values not only for the buses but also for all grid elements and their extremities, which is rarely available in real datasets; therefore, it is confined to simulations in practice.

An even more interesting factor-graph-like representation was developed in [26]. In general, a factor graph is a bipartite graph consisting of factor and variable nodes, where factor nodes represent measurement types (e.g., bus voltage and branch current), while variable nodes capture state variables (voltages). This structure allows for easy inclusion/exclusion of different measurement types and sidesteps problems with

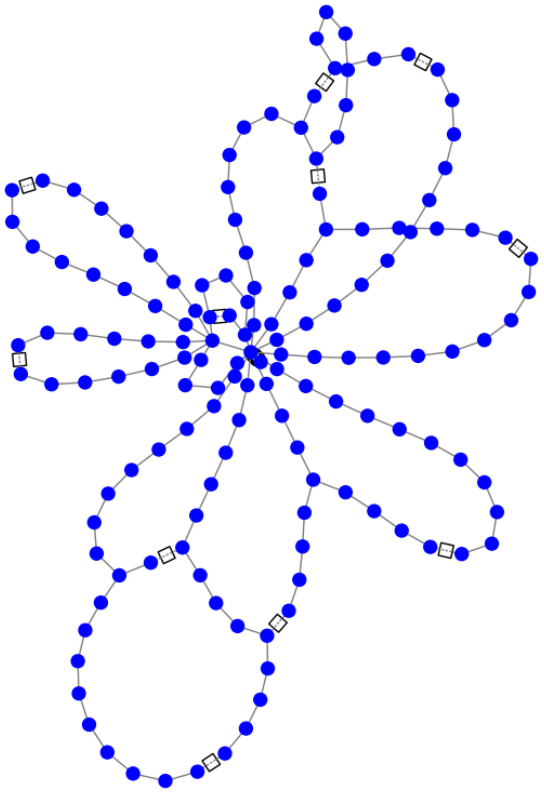


Figure 1. A visualization of the SimBench 1-MV-urban–1-sw grid.

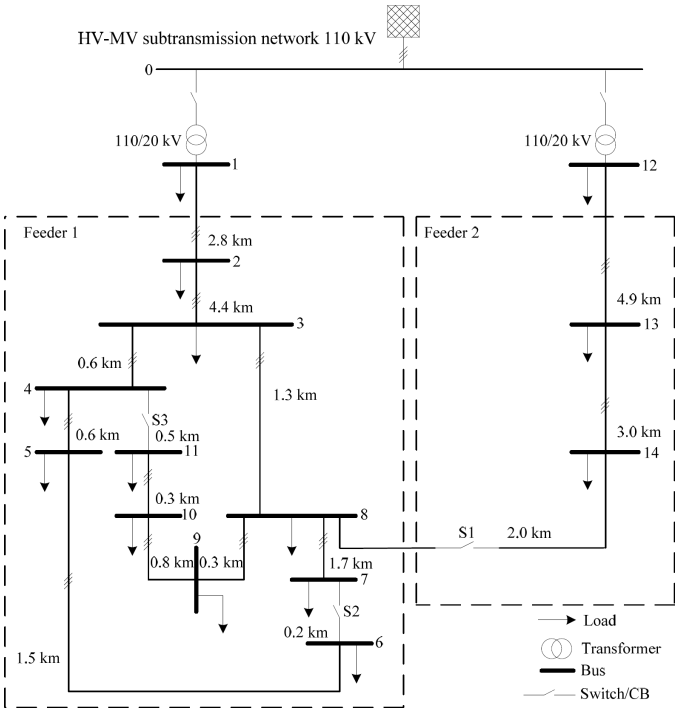


Figure 2. A visualization of the CIGRE medium voltage distribution network.

the initialization of missing features. However, it is outside the scope of our study.

E. Datasets

The main dataset used in this project is the SimBench *1-MV-urban–1-sw*, a 147-node, 10 kV medium voltage grid [27] depicted in Figure 1. It is composed of a grid model and a per-bus complex (active and reactive) power yearly time series. To calculate the resulting grid state, we performed a power flow calculation using the SIMONA energy system simulation software [28]. The resulting dataset comprises the base data and a year of complex voltage time series with a 15-minute temporal resolution. This dataset is hereafter called PQ.

Most grid branches in this model are of the open loop type, which means an open switch (depicted as a square in Figure 1) connects two separate branches. To simulate a realistic topology change, we made a line in one of the open loop branches inoperable, resembling a line fault, and closed the loop switch to resupply all nodes. Performing this operation on different branches resulted in multiple variations of the base grid topology. Afterward, we reran the simulation for each variation to obtain a topology change dataset, which is hereafter referenced as TC.

Unfortunately, the base dataset did not contain information about measurement devices. Therefore, we had to choose observable nodes randomly based on an observability level of 50%, which we assume is realistic for distribution grids. This means that the state estimator has access to true voltage values for half of the grid buses.

An auxiliary dataset used in the heterogeneous ZSL experiments is based on the CIGRE medium voltage distribution network from [29], pictured in Figure 2. It is a much smaller grid with only 15 nodes, which allows us to study how the complexity of the grids affects the performance of ZSL. The voltage data for it are generated using the Midas simulation framework [30]. The shorthand name for this dataset is MV.

V. Experiments

Our experiments are composed of three benchmarks that we call use cases. They correspond to the three subquestions of the main Research question.

Our main evaluation criterion is Median absolute deviation (MAD) across all snapshots in a dataset, which is used to estimate the average performance of models. For the ZSL experiments, we also define another metric called Performance Drop (Degradation) Ratio (PDR) as

$$\frac{MAD_{\text{train}} - MAD_{\text{test}}}{MAD_{\text{train}}}.$$

It normalizes the generalization gap, making it comparable between models or datasets. Lower PDR indicates strong generalization; higher values indicate poor generalization.

A. Static performance

Before experimenting with ZSL, we first evaluate the models without it. Here, the PQ dataset is split equally into training and testing subsets. After training on the first subset, we calculate the MAD on the second one and the PDR between

TABLE I. BEST MODEL CONFIGURATIONS FOR PQ.

Model	MAD	PDR
P2N2	0.03	-0.00
PLN	0.15	0.01
Baseline	0.36	-0.01
GIN1	0.47	-0.01
GraphSAGE1	0.53	-0.01
GraphSAGE2	0.62	-0.01
GIN2	0.63	0.00
GIN5	0.63	0.00
GraphSAGE3	0.65	-0.00
GIN6	0.67	-0.00

TABLE II. BEST MODEL CONFIGURATIONS FOR MV.

Model	MAD	PDR
P2N2	0.03	0.01
PLN	0.07	-0.00
GIN1	0.31	0.00
GraphSAGE1	0.31	0.00
GCN2	0.34	0.01
GCN3	0.35	0.00
GraphSAGE2	0.35	0.00
GAT2	0.37	0.00
GAT4	0.37	0.01
GraphSAGE10	0.37	0.00

them. The resulting values are presented in Table I. The same evaluation for the MV dataset can be found in Table II.

The near-zero PDR values for all models indicate that they can generalize to unseen data with the same topology. We can also observe that the MLP models are starting with a huge advantage, being an order of magnitude more accurate compared to GNN models, which all fall below the baseline. We acknowledge that this is not necessarily representative, as many other papers discussed in Section II are able to achieve much better performance by using different graph representations and other methods. As already mentioned, the goal of this study is not to replicate the state of the art, but rather to examine the performance changes in ZSL scenarios, which brings us to the next experiments.

As for the baseline feature propagation method, we hypothesize that it works better in higher-resolution grids where the voltage levels between nodes change more smoothly, which in our case is the PQ grid.

Comparing the performance between PQ and MV datasets, we can see that MV presents an easier task for all methods except the baseline one.

B. Observability degradation

In the first use case corresponding to subquestion A, we train the models on the grid with a baseline level of observability and then linearly reduce it to zero at testing time. Of course, the model performance decreases along with this reduction. The shorthand name of this use case is observability degradation (OD). This process is pictured in Figure 3. For readability, we limit the displayed GNN models to only the four best performing ones.

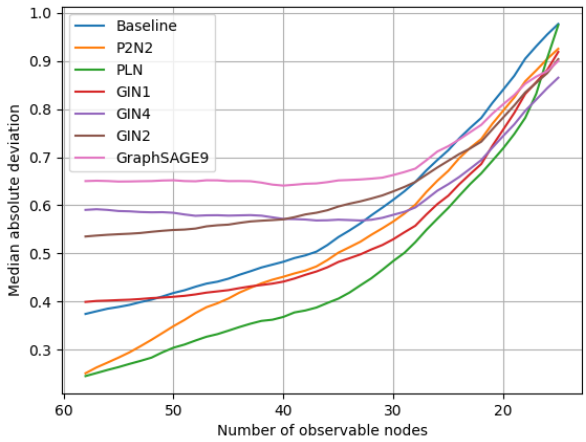


Figure 3. Performance degradation with observability reduction.

The results are unsurprising: all models behave similarly, and their performance smoothly drops from the reference values shown in Table I to the same final value at the end. GNN models maintain the reference performance longer, whereas other models lose performance immediately as the number of observable nodes decreases.

C. Homogeneous topology changes

The second use case corresponds to subquestion B and tests ZSL for homogeneous topology changes. In it, we split the TC dataset into 11 subsets according to the number of distinct topologies, meaning that the topology within each subset is static. We then perform a full 10-fold cross-validation, training the models on 10 subsets and testing on the remaining one. We then calculate the PDR for each fold. The resulting values are displayed as a box plot in Figure 4 to visualize the reliability of models in a ZSL setting.

The main observation from this graph is that the models that showed the best static performance before are now the worst performing, in terms of both averages and variation. To test if there is indeed a negative correlation, we use a scatter plot (Figure 5) of model performance on the two metrics (static MAD and median PDR in the current scenario.

The point spread suggested the existence of a Pareto front. To investigate further, we selected the non-dominated points with respect to both metrics, obtaining a bounding line that is plotted in red in Figure 5. We then computed Spearman's rank correlation on these points, which confirmed a strong negative monotonic correlation ($\rho \approx -1$, $p \approx 0$). The existence of such a clear empirical Pareto front strongly suggests a trade-off between the static accuracy of models and their robustness to homogeneous topology changes.

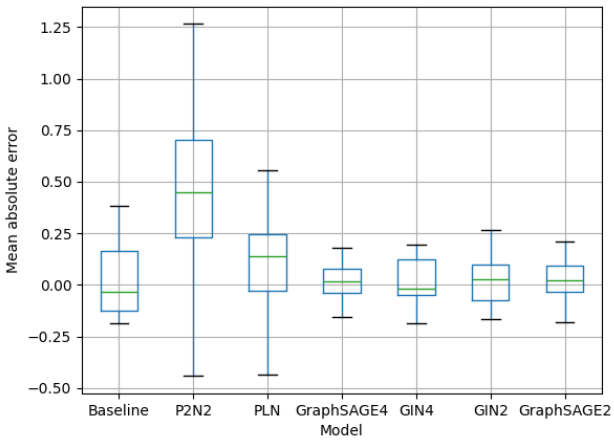


Figure 4. Cross-validation box plot.

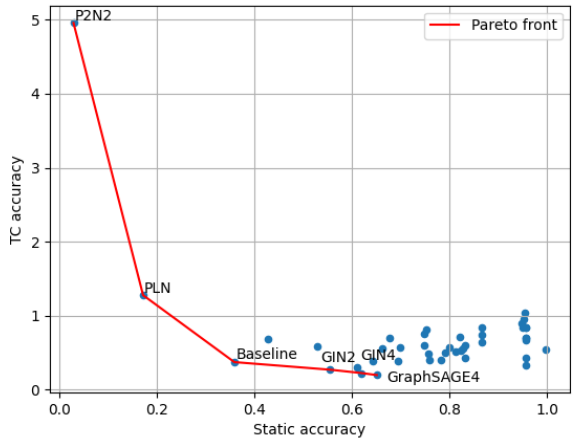


Figure 5. Cross-validation scatter plot.

The next aspect of this trade-off we wanted to analyze is the spatial error distribution that the aggregate metrics above do not capture. For this purpose, in Figure 6, we plot the comparison of the nodal error distribution between a static topology (labeled as PQ) and the switching scenario (labeled as TC). Since these plots are less space-efficient, we opted to show three cases: the MLP models (P2N2 and PLN) and an average across all GNN models, since the error distributions for these models were similar. Observable nodes are plotted as white. Note that the TC topology looks different and has a lower number of nodes due to the bus fusion transformation explained in Section IV-D.

From the figure, we can see different patterns of error propagation between the MLP and GNN models. For P2N2, errors under topology change conditions increase not just in magnitude but also in variance: the distribution of errors between nodes becomes noisy. For other models, the increase is more uniform, although certain nodes adjacent to switches pose a much harder challenge than the others. In general, failures of GNNs are less localized and instead "smeared" across the entire graph.

All models exhibit lower accuracy in long branches, which is explainable by two factors. First, switch changes mainly affect outward branches, because that is where most switches are located. Second, the outward branches are represented as path graphs, which are difficult to efficiently sample from and therefore require more sensors for robust signal reconstruction with GNNs [31].

D. Heterogeneous topology changes

The third use case corresponds to subquestion C and covers the heterogeneous ZSL scenario. Here, we transfer the model between the PQ and MV datasets in both directions, that is, training on one and then testing on another. The scenario where the model is trained on PQ and tested on MV has the shorthand name "PQ2MV", and the other "MV2PQ". We compute PDR for both transfers and present the results in the form of a scatter plot in Figure 7.

Note that it is impossible to test the MLP models in this scenario because their number of input and output features is fixed at training time, but the number of nodes between the two topologies is different. Therefore, we only test the GNN models here.

We can immediately see a significant difference between the two scenarios. Most GNNs handle PQ2MV, i.e., the transfer from a larger to a smaller network, much better than the reverse. Of course, this is in large part explainable by the fact that the MV dataset is simply an easier task, as shown above in the static performance evaluation.

A possible conjoint explanation is that the effective number of data points in a dataset for a GNN is equal to the number of snapshots multiplied by the number of nodes in the graph. Although the number of snapshots in the PQ and MV datasets is the same, the former contains more training data (and likely also more diverse data) than the latter. To test this hypothesis, we trained another series of models on a reduced PQ dataset where the number of data points is equalized with the MV dataset, and then reran the PQ2MV experiment. This increased PDR on average by 0.187, which supports the hypothesis but is not enough to fully explain the gap between PQ2MV and MV2PQ, suggesting that both factors are contributing to it.

VI. Conclusion and Future Work

Overall, the current state of the art can be represented as a three-way trade-off between conventional methods, feed-forward, and geometric models (GNNs). Conventional methods like WLS, based on the physical equations governing power systems, can provide reliably accurate results and are not affected by topology changes, as opposed to data-driven methods that infer the current state of the system from historical data. However, as we established in Section II, the amount of measurement data required for them to work is unattainable in distribution grids. But as we switch to NSE methods, we are faced with a choice between accurate feed-forward models that cannot generalize to different topologies,

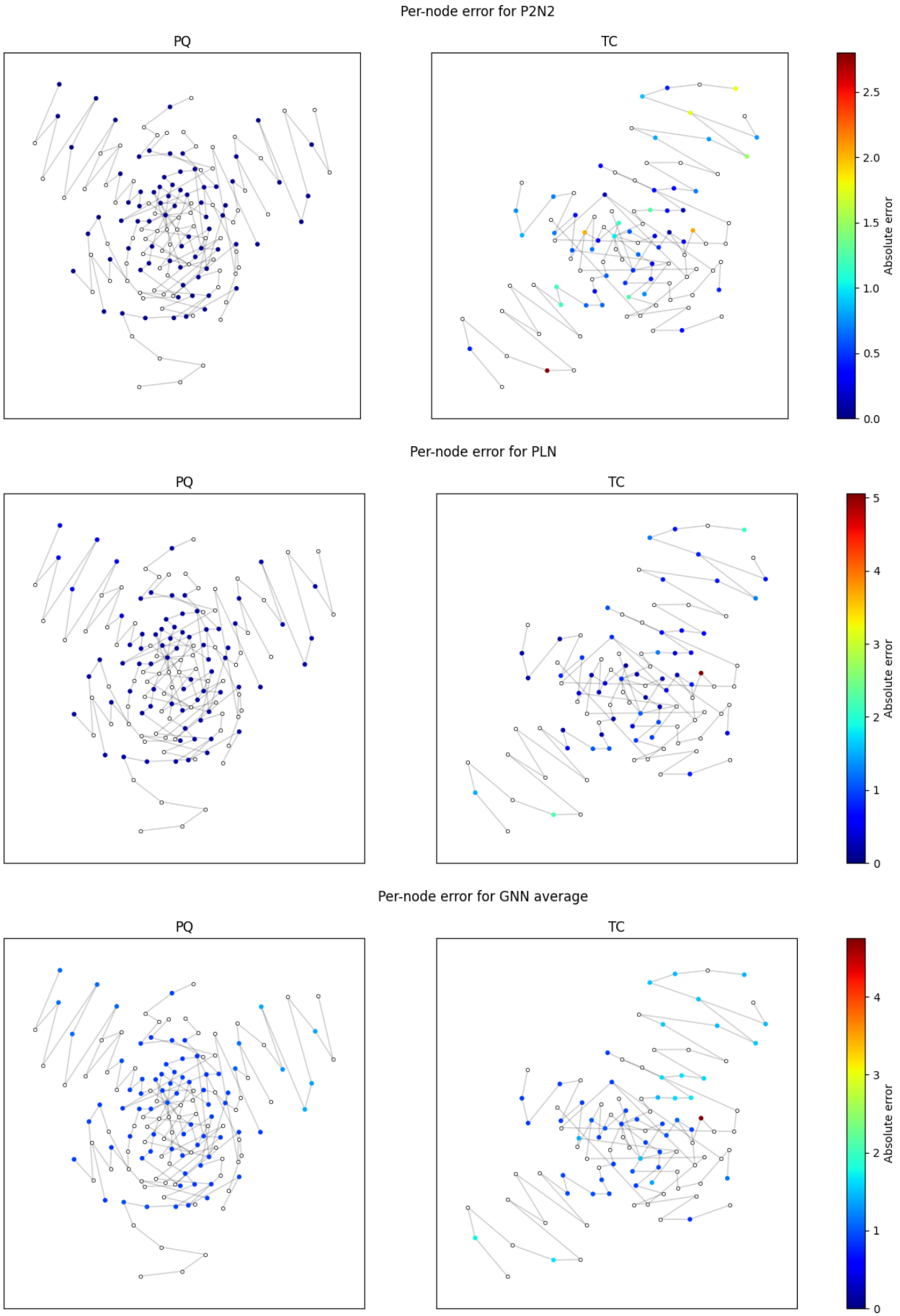


Figure 6. Nodal error comparison between static and dynamic topologies.

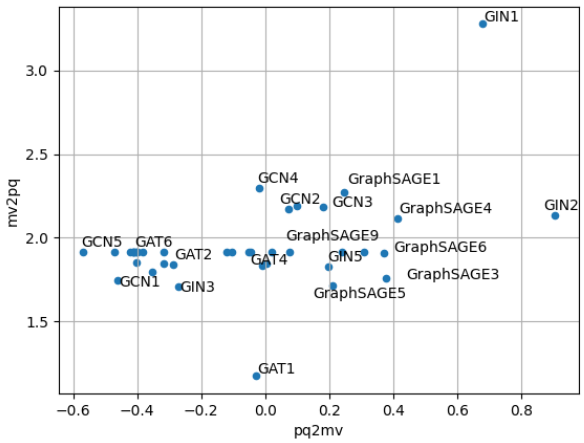


Figure 7. Heterogeneous transfer scatter plot.

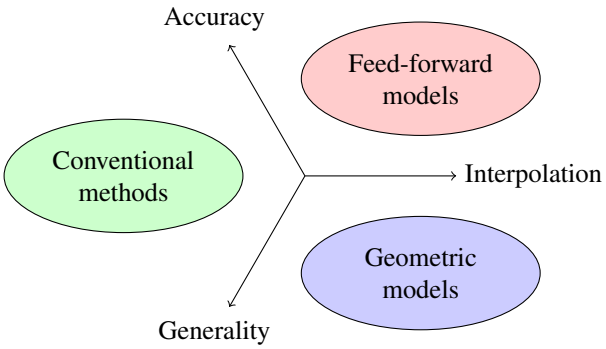


Figure 8. The three-way trade-off between conventional methods, feed-forward, and geometric models.

and GNNs, which can generalize but are not as accurate. We can therefore identify three desirable characteristics of PSSE methods:

- **Robustness to low observability**: how well does the method cope with limited measurement data, or how many sensors can be lost before the method becomes unreliable. This characteristic will be referred to as "Interpolation" in Figure 8.
- **Accuracy**: how closely can the method approximate ground truth data, assuming that there are enough data to fully utilize its expressive capacity.
- **Generality**: how well can the method adapt to changes in the grid topology without requiring retraining.

The trade-off between these characteristics is illustrated in Figure 8 and arises because no currently existing method has all three characteristics simultaneously.

Representing the problem in this way outlines the research gap: to create a method that combines performance, accuracy, and generality. This will be the direction of our future work.

Availability of Data and Source Code

The source code and datasets for this project are publicly available at the following repository:

https://gitlab.com/transense/nse-tl-paper

Acknowledgment

This research is partially supported by project TRANSENSE, funded by the German Federal Ministry for Economic Affairs and Climate Action (FKZ 03EI6044A).

References

[1] A. Berezin, S. Balduin, E. M. Veith, T. Oberließen, and S. Peter, "On zero-shot learning in neural state estimation of power distribution systems," in *ENERGY 2025, The Fifteenth International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies*, ser. IARIA Conference, Mar. 2025, pp. 47–52. [Online]. Available: https://www.thinkmind.org/library/ENERGY/ENERGY_2025/energy_2025_2_50_30034.html.

[2] F. F. Wu, "Power system state estimation: A survey," *International Journal of Electrical Power & Energy Systems*, vol. 12, no. 2, pp. 80–87, 1990, ISSN: 0142-0615. DOI: 10.1016/0142-0615(90)90003-T.

[3] A. Abur and A. G. Expósito, *Power System State Estimation*. CRC Press, Mar. 2004, ISBN: 9780203913673. DOI: 10.1201/9780203913673.

[4] K. R. Mestav, J. Luengo-Rozas, and L. Tong, "State estimation for unobservable distribution systems via deep neural networks," in *2018 IEEE Power & Energy Society General Meeting (PESGM)*, 2018, pp. 1–5. DOI: 10.1109/PESGM.2018.8586649.

[5] S. Balduin, T. Westermann, and E. Puiutta, *Evaluating different machine learning techniques as surrogate for low voltage grids*, Oct. 2020. DOI: 10.1186/s42162-020-00127-3.

[6] M. de Jong, J. Viebahn, and Y. Shapovalova, *Generalizable graph neural networks for robust power grid topology control*, 2025. arXiv: 2501.07186 [cs.LG].

[7] F. Fusco, B. Eck, R. Gormally, M. Purcell, and S. Tirupathi, *Knowledge- and data-driven services for energy systems using graph neural networks*, 2021. arXiv: 2103.07248 [cs.LG].

[8] A. Vasileiou, B. Finkelshtein, F. Geerts, R. Levie, and C. Morris, *Covered forest: Fine-grained generalization analysis of graph neural networks*, 2024. arXiv: 2412.07106 [cs.LG].

[9] S. Maskey, R. Paolino, F. Jogl, G. Kutyniok, and J. F. Lutzeyer, *Graph representational learning: When does more expressivity hurt generalization?* 2025. arXiv: 2505.11298 [cs.LG].

[10] K. Oono and T. Suzuki, "Graph neural networks exponentially lose expressive power for node classification," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=S1ldO2EFPr.

[11] X. Wu, Z. Chen, W. Wang, and A. Jadbabaie, *A non-asymptotic analysis of oversmoothing in graph neural networks*, 2023. arXiv: 2212.10701 [cs.LG].

[12] S.-G. Yang, B. J. Kim, S.-W. Son, and H. Kim, "Power-grid stability predictions using transferable machine learning.," *Chaos*, vol. 31 12, p. 123 127, 2021. DOI: 10.1063/5.0058001.

[13] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[14] T. Siameh, *Semi-supervised classification with graph convolutional networks*, Dec. 2023. DOI: 10.13140/RG.2.2.22993.71526.

[15] P. Veličković *et al.*, *Graph attention networks*, 2018. [Online]. Available: https://openreview.net/forum?id=rJXMpikCZ.

[16] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, *How powerful are graph neural networks?* 2019. [Online]. Available: https://openreview.net/forum?id=ryGs6iA5Km.

[17] W. L. Hamilton, R. Ying, and J. Leskovec, *Inductive representation learning on large graphs*, Long Beach, California, USA, 2017.

[18] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, Mar. 1964, ISSN: 0003-4851. DOI: 10.1214/aoms/1177703732.

[19] T. Cai *et al.*, *Graphnorm: A principled approach to accelerating graph neural network training*, 2021. arXiv: 2009.03294 [cs.LG].

[20] A. S. Zamzam and N. D. Sidiropoulos, "Physics-aware neural networks for distribution system state estimation," *IEEE Transactions on Power Systems*, vol. 35, no. 6, pp. 4347–4356, 2020.

[21] M.-Q. Tran, A. S. Zamzam, and P. H. Nguyen, *Enhancement of distribution system state estimation using pruned physics-aware neural networks*, 2021. DOI: 10.48550/ARXIV.2102.03893. [Online]. Available: https://arxiv.org/abs/2102.03893.

[22] L. Zhang, G. Wang, and G. B. Giannakis, "Real-time power system state estimation and forecasting via deep unrolled neural networks," *IEEE Transactions on Signal Processing*, vol. 67, no. 15, pp. 4069–4077, 2019.

[23] E. Rossi *et al.*, "On the unreasonable effectiveness of feature propagation in learning on graphs with missing node features," Proceedings of Machine Learning Research, vol. 198, B. Rieck and R. Pascanu, Eds., 11:1–11:16, Dec. 2022.

[24] L. Thurner *et al.*, "Pandapower—an open-source python tool for convenient modeling, analysis, and optimization of electric power systems," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6510–6521, 2018. DOI: 10.1109/TPWRS.2018.2829021.

[25] M. Ringsquandl *et al.*, "Power to the relational inductive bias: Graph neural networks in electrical power grids," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, ser. CIKM '21, ACM, Oct. 2021, pp. 1538–1547. DOI: 10.1145/3459637.3482464.

[26] O. Kundacina, M. Cosovic, and D. Vukobratovic, "State estimation in electric power systems leveraging graph neural networks," in *2022 17th International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, IEEE, Jun. 2022, pp. 1–6. DOI: 10.1109/pmaps53380.2022.9810559.

[27] S. Meinecke *et al.*, "Simbench—a benchmark dataset of electric power systems to compare innovative solutions based on power flow analysis," *Energies*, vol. 13, no. 12, p. 3290, Jun. 2020, ISSN: 1996-1073. DOI: 10.3390/en13123290.

[28] J. Hiry, "Agent-based discrete-event simulation environment for electric power distribution system analysis," Ph.D. dissertation, 2021. DOI: 10.17877/DE290R-22549.

[29] K. Strunz, E. Abbasi, R. Fletcher, R. Iravani, and G. Joos, *Benchmark Systems for Network Integration of Renewable and Distributed Energy Resources*. Apr. 2014, ISBN: 9782858732708.

[30] S. Balduin, E. Veith, and S. Lehnhoff, "Midas: An open-source framework for simulation-based analysis of energy systems," in *Simulation and Modeling Methodologies, Technologies and Applications*. Springer International Publishing, 2023, vol. 780, pp. 177–194, ISBN: 978-3-031-43823-3. DOI: 10.1007/978-3-031-43824-0_10.

[31] F. Wang, Y. Wang, and G. Cheung, "A-optimal sampling and robust reconstruction for graph signals via truncated neumann series," *IEEE Signal Processing Letters*, vol. 25, no. 5, pp. 680–684, May 2018, ISSN: 1558-2361. DOI: 10.1109/lsp.2018.2818062.

Power-Law Convergence in Federated Learning for Distributed Residential Load Forecasting

Alexander Wallis*, Sascha Hauke†, Hannah Jörg**, Konstantin Ziegler*
*Department of Computer Science
**Department of Interdisciplinary Studies
University of Applied Sciences Landshut
Landshut, Germany
e-mail: {alexander.wallis | hannah.joerg | konstantin.ziegler}@haw-landshut.de
†Department of Computer Sciences
Justus Liebig University Giessen
Giessen, Germany
e-mail: sascha.hauke@uni-giessen.de

***Abstract*—The integration of renewable energy resources transforms traditional energy systems, introducing prosumers entities that both produce and consume energy as key participants in modern Smart Grids. Effective load forecasting is mandatory for optimizing energy resources and grid stability. Federated Learning has emerged as a promising approach for distributed training of Machine Learning-based forecasting models. This enables collaborative model optimization across multiple prosumers while preserving data privacy. However, the impact of unbalanced data sets across participants remains a critical challenge in terms of potentially affecting learning convergence and forecast accuracy. In this work, we define and implement a Federated Learning system based on real-world electricity consumption data from a variety of prosumers. Experimental results demonstrate the trade-off between centralized and federated learning approaches, providing insights into addressing data heterogeneity in Federated Learning systems. Additionally, we show that the models convergence during training with unbalanced data sets follows a power law function. These insights highlight the potential of Federated Learning to support the evolution of distributed energy systems while ensuring data-privacy and scalability. Furthermore, the results provide actionable insights for grid operators balancing privacy, efficiency, and accuracy. Future research directions include other strategies to mitigate the effect of data imbalances and further improve the efficiency of federated optimization for dynamic energy systems.**

***Keywords*-Short-Term Load Forecasting; Federated Learning; Smart Grid; Data Privacy; Distributed Data.**

I. Introduction

This work extends the results of our conference paper [1], which provides a distributed approach for Short-Term Load Forecasting (STLF) on residential household level with respect to data privacy. Accurate load forecasting is mandatory for stable and reliable Smart Grid (SG) operation. But, the accuracy of load forecasting models, in particular Machine Learning (ML) based models, highly depends on the amount and quality of available training data [2]. Especially on smaller grid levels, e.g., low-voltage grids, or even residential household levels, the available electricity consumption data are very limited. But, with the rise of *prosumers* – consumers also able to produce electricity – prediction models on exactly this grid level are crucial for network management tasks [3].

Even if households are able to record and transmit electricity consumption data through smart meter utilization, the grid operator needs sufficient data storage and computational resources to process the data. Otherwise, the gathered data must be transferred for further processing. This transfer raises data privacy concerns and is even prohibited by law, e.g., General Data Protection Regulation [4]. The ability of information and behavior retrieval based on leakage of electricity consumption data has already been shown in the past [5], [6], [7].

Here, Federated Learning (FL) seems to be a promising approach to develop a single ML model for electricity consumption forecasting with distributed data sets – and at the same time satisfying data privacy regulation [8]. In contrast to the traditional approach, where the training of the ML model is done centralized, this task is shifted to each user individually.

In [9], FL was first used by McMahan et al. to train prediction models on mobile devices through users' keyboard inputs. Afterwards, applications with FL were proposed in various fields, e.g., medical and health care, industrial engineering, finance, transportation [10], [11], [12].

For SG development, various FL approaches were proposed, too. In [13], FL is used for anomaly detection in terms of energy usage with a detection rate compared to centralized approaches. The authors in [14] present a conceptual framework for secure FL usage in SG environments with focus on vertical and horizontal data distribution over the clients. A detailed overview of further interesting FL researches in the field of SGs is given in [15].

Although FL can be a promising approach for distributed load forecasting, the impact of unbalanced data sets among the clients is unclear. To evaluate FL in the context of prosumer-level load forecasting, we present the following contributions in this work:

- Definition and implementation of FL system composed of a variety of prosumer based on real-world electricity consumption data.

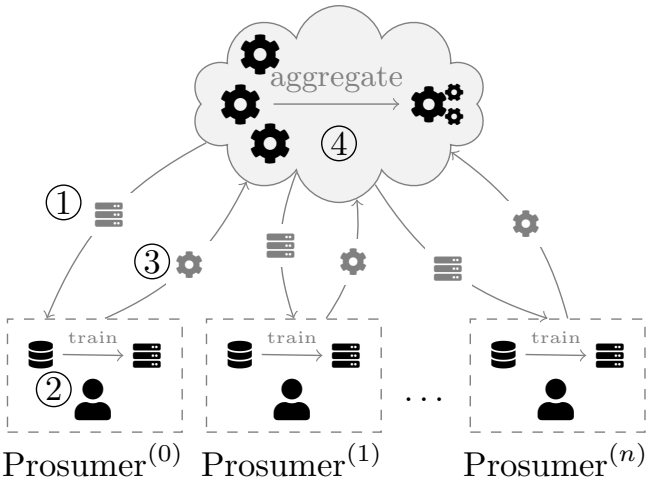


Figure 1. In a Federated Learning approach, all prosumers train their models locally on their own data.

- Comparison of forecast accuracy between a centralized and a federated learning approach for model optimization.
- Investigation of the influence of unbalanced data sets within a federation on the learning convergence and the overall forecasting error.
- Analysis of the relation between number of unbalanced clients and training convergence.

This work is organized as following. First, the necessary background information as well as notation and terminologies are given in Section II. Second, the proposed FL approach is described in detail and the different experiments conducted are described in Section III. Third, the experiment results are presented, compared, and subsequently evaluated and discussed w.r.t. forecasting accuracy in Section IV. Fourth, limitations of the proposed work and solutions are presented in Section V. Fifth and last, the insights gained from the experiments' results are summarized and starting points for further research are given in Section VI.

II. Background

Before further detailing the conducted experiments in Section III, we give the respective problem formulation (Section II-A) and background information on FL (Section II-B) as well as an overview of related work (Section II-C).

A. Problem Formulation

Basically, the load forecasting problem can be categorized into three groups based on the forecast horizon: (i) short-term, (ii) middle-term and (iii) long-term load forecasting. In this work, attention is paid on STLF, since we are interested in a household's next-day electricity consumption.

Traditionally, STLF has been addressed using both statistical and ML techniques. Early approaches include time series models such as Autoregressive Integrated Moving Average (ARIMA) and its variants [16]. With the increasing availability of high-resolution smart meter data, ML methods have gained more focus. Here, the more recent advances rely on neural networks with deep learning architectures [17]. In particular, Long-Short Term Memory Neural Network (LSTM) networks and Gated Recurrent Units (GRU) are widely used for their ability to capture temporal dependencies, whereas Convolutional Neural Networks (CNNs) and hybrid CNN-LSTM models leverage spatial and temporal features [18], [19]. In the following, the fundamental problem formulation for STLF is given.

Let $\mathbf{x}_d = (x_d^{(0)}, ..., x_d^{(T)}) \in R^T$ be a household's consumption of day d divided into T time intervals. Further, let $\mathbf{y}_d = (y_{d+1}^{(0)}, ..., y_{d+1}^{(T)}) \in R^T$ be the next day's electricity consumption, then $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i) | i = 0, ..., D\}$ is the data set composed of input-output pairs for a total of D days. Now, a supervised learning approach approximates a function $\mathbf{y}_d \approx \hat{f}(\mathbf{x}_d)$ for the following optimization problem:

$$\arg\min_{\hat{f} \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} L(\hat{f}(\mathbf{x_i}), \mathbf{y}_i) \tag{1}$$

where $L(\cdot)$ is the desired cost function to be minimized.

Typically, in a centralized learning setting, this is done by collecting each household's data and subsequently by training a combined forecasting model, which is afterwards distributed to every household. Indeed, this rises all of the problems and concerns described earlier (see Section I) and FL is a promising approach to tackle all of them.

B. Federated Learning

Contrary to the centralized learning, a FL approach guarantees data-privacy by preserving prosumers' consumption data locally. A collaboration of prosumer – a so-called *federation* – trains a STLF model by only exchanging respective model parameters. Typically, the participants within a federation are called clients but in this work the terms clients, prosumers and households are used interchangeably. Let $\mathcal{P} = \{p^{(i)} | i = 0, ..., N\}$ be the set of N prosumers then FL procedure involves the following steps:

1) **Distribution** of the initial global model to all prosumers which are part of the federation $p \in \mathcal{P}$.
2) **Training** of the global model by adjusting it's parameters based on the local data set of every prosumer.
3) **Returning** the adjusted model parameters to a central unit, e.g., trusted 3rd party, data center, one of the participants.
4) **Aggregation** of all received parameters by a predefined `aggregate`-function and integration into the global model.

This whole procedure, also depicted in Figure 1, is repeated over a defined number of *communication rounds* r. Interestingly, reducing the number C of clients participating in every learning round increases the communication efficiency without loss of prediction accuracy [9]. So, in every round a prosumer subset $\mathcal{P}'_r \subseteq \mathcal{P}$ with $|\mathcal{P}'_r| = C$ is randomly chosen to take part in the training task in step 2.

Beside the number of prosumers involved in training, the used `aggregate`-function offers additional flexibility. In [9],

the author introduces FedSGD and FedAvg, where the later is the common approach for solving the FL problem by calculating the (weighted) average (often mean) per parameter. Other aggregation approaches are, e.g., federated adaptive optimizers (FedAdam, FedAdagrad, FedYogi) [20], momentum-based variance-reduced technique (FAFED) [21], heterogeneity focused (FedProx [22], SCAFFOLD [23]). There are plenty more proposed aggregate-methods, and the related questions in terms of, e.g., applicability, optimality, generalization, are major research topics.

At this point, it is worth noting that additional security mechanism are needed to guarantee some desired security level. Although, FL offers a framework for data-privacy in distributed learning, data leakage or reconstruction attacks are still possible [24]. Privacy enhancing techniques applicable for FL settings are, e.g., differential privacy and homomorphic encryption [25].

In the next section, we give an overview of existing FL research with focus on STLF.

C. Related Work

After describing the FL approach in general, we give an overview of existing FL research conducted in the field of residential STLF. Here, we limit the related work explicitly to (i) residential households and (ii) maximum 24-hour forecast horizon.

A comparison between FedAvg and FedSGD with different forecast horizons (1 h and 24 h) is given in [26]. They showed that their proposed FL model with FedAvg reaches higher accuracy than a centralized and a personalized model.

In [27], the authors compare the forecasting accuracy of a FL model on prosumers involved in training and on hold-out prosumers. They choose this approach to evaluate how well the global model fit for non-participating prosumers. Here, the non-participant prosumers fine tune the pre-trained model for 5 epochs locally. They conclude that this fine tuning step improves the forecast accuracy compared to the global model.

In terms of unbalanced client data distribution, Liu et al. proposed the closest approach [28]. Here, clients are divided into 5 groups based on the resolution of their available consumption data ranging from 300 s to 1.800 s.

A hybrid CNN-LSTM model is used in a FL setting in [29]. To handle the consumption heterogeneity, the authors propose a model fine-tuning step after the weight aggregation based on multiple kernel variant of maximum mean discrepancies. Furthermore, all clients are involved in every training and the number of data samples are equal over all clients.

The authors in [30] compare the accuracy of a centralized model with a FL one, a FL plus clustering, and FL plus clustering and subsequently local fine tuning. Here, the last approach reaches the highest accuracy. But, to manage all experiment permutations the evaluations are done with fixed $C = 0.1$.

A personalized FL approach is presented by Rahman et al. [31]. Here, a meta-learning-based strategy is applied such that each client trains their local LSTM with different learning rates. This strategy is developed to address data heterogeneity among the clients. The provided simulations show that their personalized approach reach higher prediction accuracy than traditional LSTM as well as FL approaches.

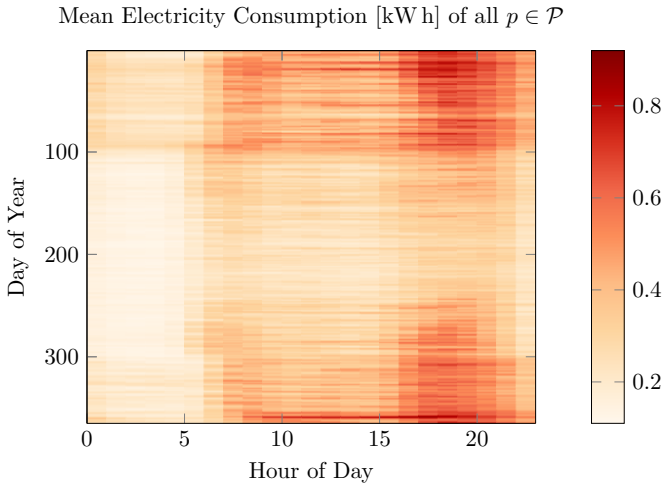


Figure 2. Mean electricity consumption of all selected households from the SmartMeterInLondon data set.

All of the mentioned related work are summarized with their respective training and model parameters in Table I. It can be seen that the related work in terms of unbalanced data sets is non existing – as far as we know – for the STLF problem on residential prosumer level.

III. Methodology

To evaluate our proposed FL approach, different experiments are conducted in this work. Therefore, we build a federation composed of prosumers represented by household data taken from public available real-world electricity records (see Section III-A).

A. Used Data Set

In this work, residential household data are taken from the SmartMetersInLondon [32] data set, which is a refactored version of the "Low Carbon London Project" data. This data set contains electricity consumption records for 5,567 London households between November 2011 and February 2014. In the following, the conducted data preprocessing and preparation steps as well as the selection of suitable households is described.

a) Household Selection: Since the date range differs between prosumers in the data set, only houses with the most overlap are selected. Furthermore, households with more than three consecutive hours of missing values are removed – otherwise, missing values are linearly interpolated. In total, 20 households are selected suitable for further usage. The hourly mean electricity consumption is depicted for every day in the training set in Figure 2. Subsequently, the respective consumption data is preprocessed for every selected household in the following.

TABLE I. OVERVIEW AND SUMMARY OF RELATED WORK FOR FEDERATED LEARNING (FL) APPROACHES FOR RESIDENTIAL SHORT-TERM LOAD FORECASTING (STLF).

Related Work	#Clients	C	ML-Model	Data Set	Balanced Data	Aggregation
Taïk and Cherkaoui [27]	200	$5, 10$	LSTM	AUSTIN	yes	FedAVG
Fekri et al. [26]	19	6	LSTM	non-public	yes	FedSDG, FedAVG
Liu et al. [28]	50	10	iQGRU	AUSTIN	semi	FedAVG
Shi and Xu [29]	10	10	CNN-LSTM	LONDON	yes	FedAVG
Briggs et al. [30]	100	0.1	LSTM	LONDON	yes	FedAVG
Rahman et al. [31]	5	5	LSTM	FRANCE	both	FedAVG
our work [1]	20	$1, 2, 5, 10, 20$	MLP	LONDON	no	FedAVG

b) Data Preprocessing: Since the date ranges of available data varies tremendously across all prosumers, we select the time between 1st January 2013 and 28th February 2014 with the most overlapping data. This interval is further divided into train and test data ($\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$), whereas the whole year 2013 is used for training and the remaining data for testing. This leads to $|\mathcal{D}^{(p)}_{\text{train}}| = 8,760$ and $|\mathcal{D}^{(p)}_{\text{test}}| = 1,416$ samples for every prosumer. For every prosumer, both data sets are rescaled individually with the standardization given by

$$x' = \frac{x - \sigma}{\mu}, \tag{2}$$

where x' is the transformed consumption time series with mean (μ) of 0 and standard deviation (σ) of 1 (unit variance).

c) Look-back and Forecast Horizon: The accuracy of time series forecasting depends on both, the chosen look-back window as well as the forecast horizon. In the related work (Section II-C), those parameter differ across studies. Here, our proposed forecasting model uses the last 24 h as input to predict the next 24 h. Although, additional features, e.g., weather, holiday, weekday/weekend, can reduce the forecast error, we restrict our model to the raw consumption values. In [33], we evaluate the FL model with further feature engineering. So this leads to an input vector $\mathbf{x} = (x_t, x_{t-1}, ..., x_{t-23}) \in \mathbb{R}^{24}$ and an output vector $\mathbf{y} = (y_{t+1}, y_{t+2}, \ldots, y_{t+24}) \in \mathbb{R}^{24}$ for every day and for each prosumer in the data set.

After the household selection and necessary preprocessing steps, the used ML model architecture, as well as further details on the overall development process is given in the next part.

B. System Setting

In this section, we give all the relevant information about the model architecture and used hyperparameters. Afterwards, a definition for different kinds of learning prosumers within the federation based on the ability to store training data is presented. A description of the used federation, as well as the training procedure is given in the third part.

a) Model and Hyperparameters: In this work, we choose a vanilla Multi-Layer Perceptron (MLP) as model architecture, similar to the proposed model in [9]. This architecture allows an easy implementation and training on lightweight devices with limited computational resources. This fully connected MLP has two hidden layers with 200 neurons each and uses a Rectified Linear Unit (ReLU) as activation function.

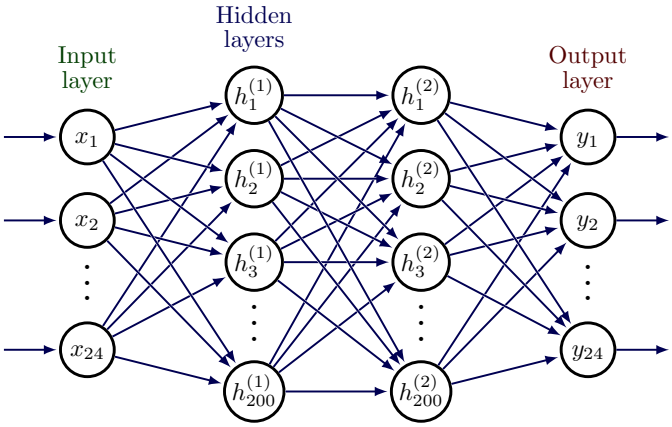


Figure 3. Network architecture used in this work. Fully connected MLP with 2 hidden layers, 200 neurons each and ReLU activation function. The input $\mathbf{x} = (x_{t-24}, \ldots, x_t)$

$$\text{ReLU} = \max(0, x) = \begin{cases} x \text{ if } x > 0 \\ 0x \leq 0. \end{cases} \tag{3}$$

The final model architecture used in this work for every experiment is illustrated in Figure 3.

b) Weak and Strong Prosumer: We introduce the terms *strong* and *weak* prosumer, to describe two different types of prosumers based on the amount of available training data. The two types are defined the following way:

Definition 1. Let $p \in \mathcal{P}$ be a prosumer only able to store training data between two consecutive communication rounds, then it is called a *weak* prosumer p_{weak}.

Definition 2. Let $p \in \mathcal{P}$ be a prosumer with no storage limitations, then it is called a *strong* prosumer p_{strong}.

Based on the Definitions 1 and 2, we define the fraction of strong prosumers within a federation as the so-called *strong-prosumer-fraction*:

Definition 3. Let $|p_{\text{weak}}|, |p_{\text{strong}}|$ be the number of weak respective strong prosumers in $\mathcal{P}$, then the strong-prosumer-fraction is defined as $\phi = \frac{|p_{\text{strong}}|}{|p_{\text{weak}}|+|p_{\text{strong}}|}$.

This allows a straightforward distinction between prosumers within a federation and introduces another parameter for the overall training procedure.

c) Training Procedure: For all conducted experiments, with or without strong and weak prosumers, the respective training procedure takes $r = 100$ communication rounds in total. At $r = 0$ the global model's weights w are randomly initialized. After every round, the global model's weights are updated by a weighted `FedAvg` aggregation function, given as

$$w_{r+1} \leftarrow \sum_{p \in \mathcal{P}'_r} \frac{n_p}{n} w_r^{(p)}, \tag{4}$$

where n_p, n is the number of sample per prosumer respective the number of all samples. The local weights $w_r^{(p)}$ are calculated locally for every $p \in \mathcal{P}'_r$ in parallel by

$$w_r^{(p)} \leftarrow w_r - \eta \nabla_w \mathcal{L}(w_r; \mathbf{x}_i, \mathbf{y}_i) \tag{5}$$

for a single epoch with a learning rate of $\eta = 0.001$ and the Mean Squared Error (MSE) as loss function $\mathcal{L}(\cdot)$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2, \tag{6}$$

where n is the number of test set samples and $\hat{y}_i, y_i$ is the predicted respective actual consumption value.

To evaluate the proposed FL approach and also to analyze the impact of unbalanced data sets, various experiments are conducted, which are further detailed in the following section.

C. Experiment Settings

The proposed FL approach for residential STLF is evaluated in different experiments. The evaluation is based on the MSE error metric given in Equation 6. In total, we run the following three experiments:

I **Benchmark** A centralized model – as well as one local model for every prosumer – is trained over r epochs.

II **Number of Learners** Since a new subset of learning prosumers is selected in every round (see Section II-B), we evaluate the model's forecast accuracy for different number of learners $C = \{1, 3, 5, 7, 10, 20\}$.

III **Strong Prosumer Fraction** With the introduction of weak and strong prosumers, we evaluate our FL approach based on unbalanced data sets. For $C = \{1, 10, 20\}$ the strong-prosumer-fraction $\phi = \{0.05, 0.25, 0.5, 0.75, 1\}$ is considered. Here, the unbalanced data set evolves over the communication rounds $r = \{1, 2, \ldots, 100\}$ by:

$$\text{weak:} \quad \mathcal{D}_r^{(p)} = \mathcal{D}_{r-1:r}^{(p)} \tag{7}$$

$$\text{strong:} \quad \mathcal{D}_r^{(p)} = \mathcal{D}_{0:r}^{(p)}. \tag{8}$$

So, for strong prosumer the training samples increase by $n = \lfloor \frac{|\mathcal{D}|}{r} \rfloor$ in every round, whereas for weak prosumer the samples have a fixed size of n.

The experiments I-III are repeated for $N = 10$ times to handle the randomness via model initialization and prosumer sampling with C, ϕ. Our proposed FL approach is implemented in `Python=3.9` with `PyTorch` and model training was executed on a local machine with a Nvidia Geforce RTX 2080 graphic card. The experiments' results are listed in the next section.

TABLE II. TEST SET ERROR FOR EXPERIMENT I. MSE IS CALCULATED OVER ALL 20 PROSUMERS

Model	↓ MSE ($\mu \pm \sigma$)	min	max	won
centralized	0.181 ± 0.13	0.030	0.545	3 out of 20
personalized	0.166 ± 0.13	0.021	0.514	17 out of 20

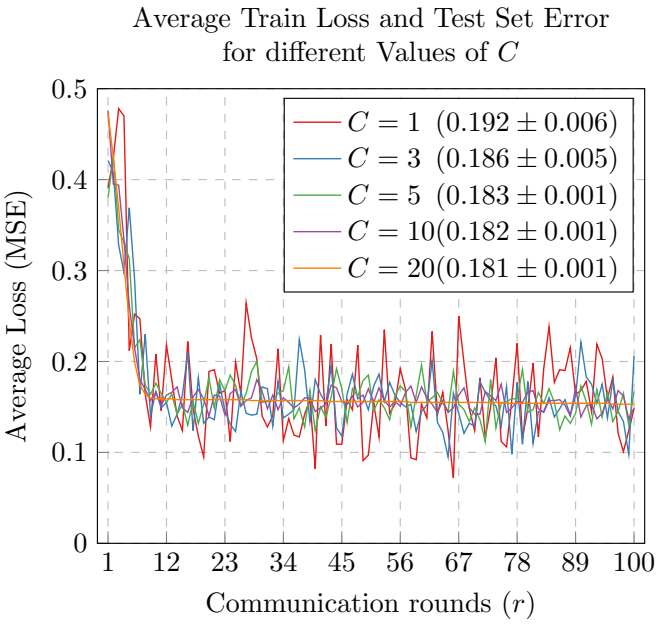


Figure 4. Experiment II: Train loss and test set error with mean and standard deviation over 10 repetitions for different values of C.

IV. EXPERIMENT RESULTS & DISCUSSION

The results of the various experiments are presented in the same order as defined in Section III-C. The respective results are provided below, followed by a detailed analysis and discussion.

Figure 4 illustrates the training loss across all communication rounds r as well as the test set error in the legend. For the different values of $C = \{1, 3, 5, 10, 20\}$, the test set error is given as mean with standard deviation over all 10 repetitions. Similar to experiment I, the MSE is calculated over all prosumers $p \in \mathcal{P}$ without individual examination.

For experiment III, results are given in two ways. First, the average training loss over all runs is depicted in Figure 5. Second, Table III lists the test set errors. In addition to numerical values over all prosumers, the MSE is also calculated separately for the sets of p_{weak} and p_{strong}. The minimum and maximum MSE values are determined over all 10 runs combined for each combination of C- and ϕ-values.

In this work, a FL approach was proposed for the STLF problem at residential prosumer level. Three experiments were

TABLE III. TEST SET ERROR FOR EXPERIMENT III. ERROR IS GIVEN AS MSE WITH MEAN AND STANDARD DEVIATION OVER ALL 10 REPETITIONS.

C	ϕ	↓ MSE ($\mu \pm \sigma$)				
		all	strong	weak	min	max
1	0.05	*0.215 ± 0.14*	0.192 ± 0.12	0.216 ± 0.14	0.026	0.674
	0.25	**0.193 ± 0.12**	0.209 ± 0.15	0.188 ± 0.11	0.039	0.597
	0.5	0.194 ± 0.12	0.202 ± 0.13	0.186 ± 0.12	0.037	0.565
	0.75	0.196 ± 0.12	0.194 ± 0.12	0.199 ± 0.13	0.038	0.587
	1	0.201 ± 0.13	0.201 ± 0.13	–	0.036	0.626
10	1	*0.223 ± 0.15*	0.142 ± 0.07	0.227 ± 0.15	0.029	0.750
	0.25	0.186 ± 0.12	0.187 ± 0.13	0.186 ± 0.11	0.033	0.540
	0.5	0.184 ± 0.12	0.182 ± 0.12	0.187 ± 0.12	0.038	0.550
	0.75	0.181 ± 0.12	0.185 ± 0.12	0.170 ± 0.10	0.038	0.525
	1	**0.180 ± 0.12**	0.180 ± 0.12	–	0.041	0.527
20	1	*0.198 ± 0.13*	0.205 ± 0.13	0.198 ± 0.13	0.034	0.711
	0.25	0.190 ± 0.12	0.193 ± 0.13	0.189 ± 0.12	0.035	0.591
	0.5	0.183 ± 0.12	0.173 ± 0.11	0.192 ± 0.13	0.040	0.546
	0.75	0.181 ± 0.12	0.172 ± 0.11	0.208 ± 0.12	0.038	0.523
	1	**0.179 ± 0.11**	0.179 ± 0.11	–	0.042	0.516

Note: lowest error is in **bold**, highest in *italic*.

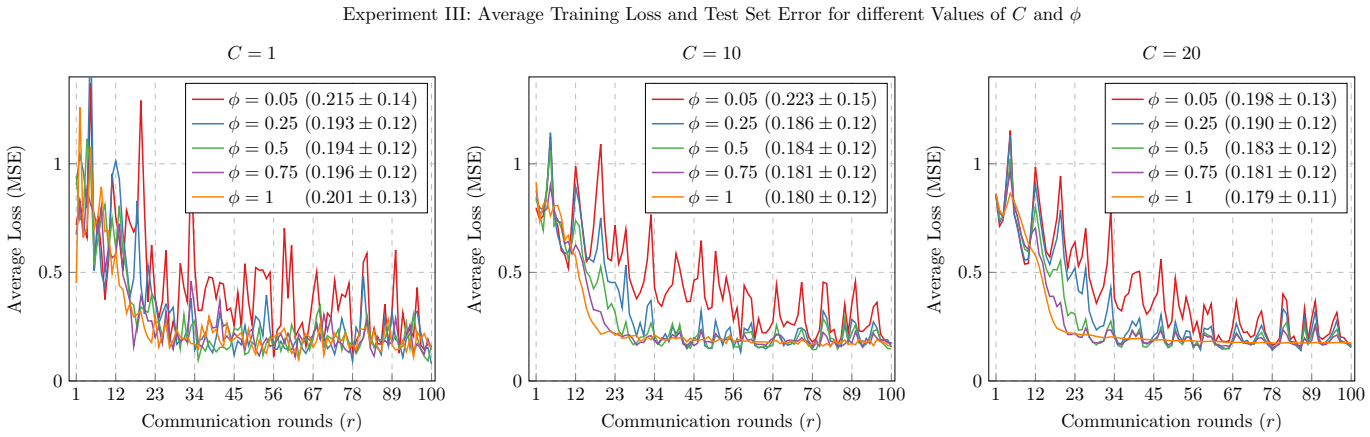


Figure 5. The training loss and test set error for different fractions of strong prosumer ϕ evaluated for $C = 1$ (left), $C = 10$ (middle), and $C = 20$ (right).

conducted to analyze the impact of unbalanced data distribution among prosumers within the federation.

The first experiment compared a centralized MLP trained on all prosumers' data with a personalized MLP trained individually for each prosumer. Of 20 households in total, 17 times the personalized model reaches a higher accuracy (see Table II). This indicates a strong distribution of consumption behaviour across the prosumers since more data does not guarantee better results.

The second experiment examined the effect of different numbers of learners. As shown in Figure 4, test set errors show minimal variation for $C > 1$, with nearly identical training loss reduction. However, lower C-values introduce more variance, emphasizing trade-off between distribution computational resources and learning efficiency.

In real-world scenarios, training data availability varies among prosumers due to recording and storage capabilities as well as temporal offsets in joining the federation. To address this, the third experiment introduced the distinction between weak and strong prosumers, defined by storage capability.

TABLE IV. SUMMARY OF COMMUNICATION ROUNDS TO REACH TARGET MSE FOR 10 RUNS WITH MEAN (μ) AND STANDARD DEVIATION (σ).

C	ϕ	Rounds to target MSE			
		Mean (μ)	STD (σ)	Min	Max
10	0.05	64	14	48	84
	0.25	31	6	24	42
	0.5	24	2	22	27
	0.75	21	3	16	26
	1	18	2	15	22
20	0.05	60	16	48	98
	0.25	34	8	26	55
	0.5	26	2	20	28
	0.75	22	2	20	26
	1	31	2	29	35

The strong prosumer fraction ϕ represents the proportion of strong prosumers within a federation. Figure 5 indicates slower training convergence with a decreasing number of strong prosumers, irrespective of C-values. However, reducing ϕ to 0.75 or 0.5 did not significantly impact training speed or test set error. This finding is relevant for practical applications, suggesting that not all prosumers need to contribute learning resources to maintain overall performance. This is further discussed in the following part.

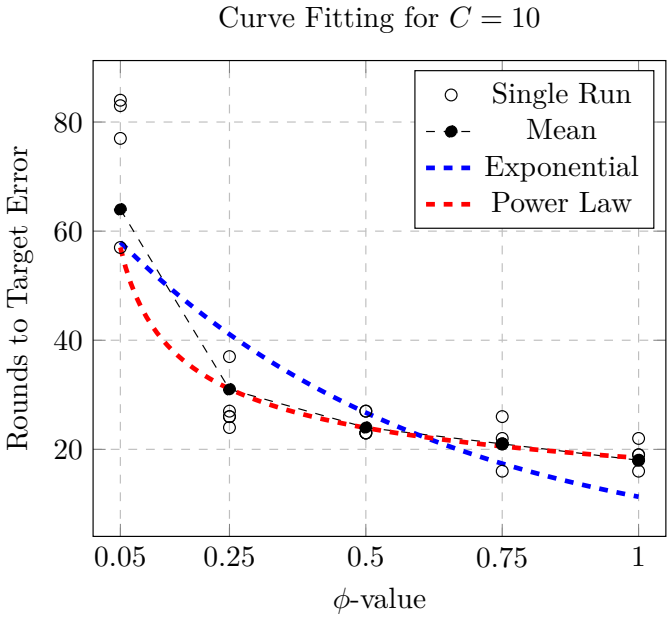
A. Impact of Strong Prosumer Fraction

To evaluate the impact of stron prosumer fraction ϕ on the overall training convergence, we introduce a target error $\text{MSE}_{\text{target}} = 0.20$. This value is chosen based on the test set error from Table III, where all lowest errors are below this threshold. Afterwards, we determine the communication rounds needed to reach the desired train error $\text{MSE}_{\text{train}} < \text{MSE}_{\text{target}}$. This is repeated for all ϕ-values and every single run. The mean (μ) numbers of needed communication rounds are listed in Table IV with respective standard deviations (σ). Where, we only consider $C = 10$ and $C = 20$ since for the $C = 1$ case, the variance over all runs is too high to get meaningful results.

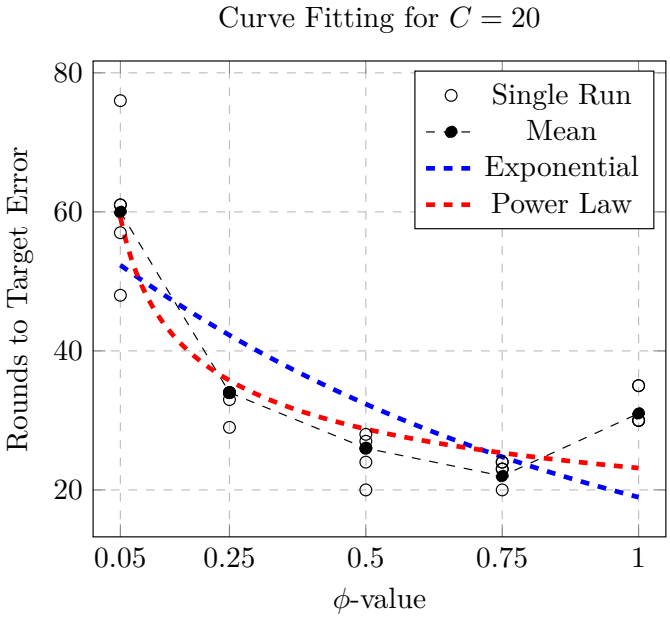
From Table IV, we can see that with increasing ϕ-value the number of communication rounds to reach the target MSE is decreasing. This finding is also the case for both experiments with $C = 10$ and $C = 20$ which indicates some degree of relation between needed communication rounds and amount of strong prosumers within the federation. This relation is further analyzed in the following part.

B. Curve Fitting

The rounds per ϕ-value to reach the target MSE from Table IV are shown in Figures 6a and 6b for $C = 10$ respective $C = 20$. In this figure, the decreasing trend with increased ϕ-value is clearly recognizable. Furthermore, it seems that the first few additional strong prosumers lead to the highest reduction



(a)



(b)

Figure 6. Curve fitting results for (a) $C = 10$ and (b) $C = 20$ with R^2-score metrics for exponential (red) and power law (blue) functions.

in communication rounds and therefore a non-linear relation is possible. In the following, we examine two feasible functions, namely exponential and power law, defined as:

$$r_{\text{exp}} = a * \exp(\phi * b), \tag{9}$$

$$r_{\text{pow}} = a * \phi^b. \tag{10}$$

Here, the dependent variable ϕ is the strong prosumer fraction and the independent variable r is the number of communication rounds. To estimate the functions' parameters

TABLE V. GOODNESS OF FIT METRICS FOR THE CURVE FITTED MODELS: EXPONENTIAL AND POWER LAW.

C	Model	$\uparrow R^2$	$\downarrow$ RMSE	$\downarrow$ AIC	$\downarrow$ BIC
10	exponential	0.86	6.38	22.53	21.75
	power law	0.99	0.65	-0.24	-1.02
20	exponential	0.64	7.99	24.79	24.01
	power law	0.91	4.09	18.12	17.33

$\Theta = [a, b]$, we employ a curve fitting based on non-linear least squares fitting approach.

$$\hat{r}_{\text{exp},10} = 63.179 \cdot \exp^{(-0.086 \cdot \phi)} \tag{11}$$

$$\hat{r}_{\text{pow},10} = 57.005 \cdot \phi^{-0.377} \tag{12}$$

$$\hat{r}_{\text{exp},20} = 55.216 \cdot \exp^{(-0.053 \cdot \phi)} \tag{13}$$

$$\hat{r}_{\text{pow},20} = 59.196 \cdot \phi^{-0.313}. \tag{14}$$

Given the observed data from Table II, the fitting process estimates a parameter vector Θ^* that maximizes the sum of squared residuals

$$\Theta^* = \arg\min_{\Theta} \sum_{i=1}^{N} (r_i - f(\phi, \Theta))^2. \tag{15}$$

This minimization problem is solved using the `Levenberg-Marquardt` algorithm [34], [35], which is suitable for small- to medium-sized problems with smooth, differentiable models. After solving the optimization problem, the following functions are estimated

The curve fitting is performed with the Python package `scipy.optimize`. In Figure 6, the fitted functions are shown in dashed red (power law) and blue (exponential) lines. Figure 6 suggests that a power law function describes the data points more accurate than the exponential function. But to quantify the results, *goodness of fit* metrics are applied in the following part.

C. Goodness of Fit

To evaluate the performance of the fitted models and their abilities to explain the observed data (see Table IV), we assess the goodness of fit using the *coefficient of determination*, known as R^2 metric. This metric provides a normalized measure of how much the total variance in the observed data is accounted by the respective model. Originally, the R^2 metric was developed for linear regression models, but it is widely used and applicable in non-linear context as an indicative summary statistic [36]. Formally, the R^2-score is defined as:

$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\sum_{i=1}^{n}(r_i - \hat{r}_i)}{\sum_{i=1}^{n}(r_i - \bar{r}_i)}, \tag{16}$$

where SSR is the residual sum of squares and SST is the total sum of squares. Thus, a R^2 score of 1 indicates a perfect fit. Conversely, an R^2 score of 0 indicates that the model performance is worse than a predicting the mean of the observed data.

Although, R^2 is useful for summarizing model fit, it should not be the only metric for model evaluation, especially in non-linear settings. Therefore, we further calculate the Root Mean Squared Error (RMSE) given as:

$$\text{RMSE} = \sqrt{\text{MSE}}, \tag{17}$$

and the Akaike Information Criterion (AIC) as well as the Bayesian Information Criterion (BIC):

$$\text{AIC} = n \cdot \log \frac{\text{SSR}}{n} + 2k \tag{18}$$

$$\text{BIC} = n \cdot \log \frac{\text{SSR}}{n} + k \cdot \log n, \tag{19}$$

where k is the number of parameters and n the number of data points. The various metrics are listed in Table V. Based on the provided metrics, the power law models are able to describe the observed data more precise than the exponential ones. The implications are discussed in the next section.

D. Curve Fitting Implications

The observation that a power law model provides a better fit to the data than an exponential model carries important implications about the underlying system dynamics. In this case, the highest decrease in communication rounds happens within the first few additional strong prosumers. All further additions have only diminishing returns. This insight is of great interest for power grid operators and the development of distributed smart micro grids since only a few strong clients, e.g., households, are enough to accelerate training duration and therefore improve forecast accuracy for all clients within the federation.

V. LIMITATIONS

While experiment results highlight the potential of FL for the STLF problem at residential household level, several limitations need to be acknowledged. First, the provided study relied on a MLP neural network architecture with `FedAvg` as aggregation method. Although, this was a planned choice to ensure comparability with prior work and to enable implementation on distributed micro computers, it excludes more advanced model architectures, e.g., Long-Short Term Memory Neural Network (LSTM), Transformer-based models, GRU, and aggregation strategies, e.g., `FedProx`, `FedAdam`, which could yield higher forecasting accuracy. Future work should validate whether the observed power law convergence persists across those architectures. Second, the experiments were conducted with a single data set (`SmartMetersInLondon`, see Section III-A). While this data set is publicly available and also provides sufficient diversity across multiple households, it is limited to a specific geographic, temporal, and regulatory setting. Other regions may reveal different consumption patterns. Therefore, the generalization to rural grids, microgrids, or regions with higher renewable energy resources remains uncertain. Third, our proposed model restricted the input space to past consumption data without including exogenous features as weather data, calendar effects, or socio-economic indicators (see Section III-A). While the experiment design focused on

unbalanced data distribution among the clients within the federation, the true forecasting potential of FL models may not be exhausted.

Despite these limitations and constraints, the findings provided by this study hold relevant implications for both research and practical application. For grid operators, the observation that only a small fraction of strong prosumers is necessary to accelerate convergence suggests that FL can be made efficient without universal data in high-resolution. This reduces infrastructure requirements and communication overhead. For prosumers, FL provides a possibility to contribute to a forecasting model without disclosing privacy-sensitive consumption data, which aligns with regulations such as the GDPR. In summary, while the presented experiments have clear methodological boundaries, they provide valuable evidence that FL can balance accuracy, efficiency, and privacy in real-world smart grid environments. The listed limitations also provide promising directions for advancing future research areas.

VI. Conclusion & Future Work

This work developed a ML-based model for the STLF problem at residential prosumer level. Given that high-resolution electricity consumption data contain behavioral information, data privacy concerns arise when transferring and processing such data. To address this, FL was incorporated as a viable approach to train ML models on distributed data without requiring direct data exchange. Three experiments were designed and conducted to evaluate the proposed FL approach. The results demonstrated that FL can achieve competitive forecasting accuracy while preserving data privacy. The trade-off between the number of learners and computational efficiency was also analyzed, along with the effects of strong and weak prosumers on training convergence and performance. Additionally, limitations of our provided work are discussed and possible solutions in future work are given.

In future work, we will focus on extending and improving the proposed FL approach. This study primarily addressed unbalanced data sets within a federation, adopting constraints such as a lightweight MLP architecture, state-of-the-art `FedAvg` weight aggregation, and the exclusion of external features. To enhance overall forecasting accuracy, these constraints should be revisited. Preliminary results indicate the utilizing more complex LSTM models and incorporating weather information can reduce forecasting errors. Additionally, this study did not explicitly implement a security layer. Future research will explore methods to ensure data privacy and prevent information leakage while integrating insights from this study. Furthermore, the potential of Transformer-based models for STLF remains an unexplored area, warranting future investigation.

Additionally, future research could explore the integration of transfer learning techniques, where forecasting knowledge gained in one region or community is transferred to another. This allows FL models trained on areas with sufficient data to support rural or emerging smart grid regions. Another promising direction is the study of incentive mechanisms for prosumers. Since FL requires active participation, especially from strong prosumers, future work should consider incentives that reward households for contributing computational resources and data.

Acknowledgment

The research in this work is supported for Alexander Wallis by the Federal Ministry of Education and Research BMBF (FKZ 03FHP212) and for Sascha Hauke by Bayern Innovativ. This is an extended version of our work presented at the *15th International Conference on Smart Grids, Green Communication and IT Energy-aware Technologies*

References

[1] A. Wallis, S. Hauke, H. Jörg, and K. Ziegler, "Federated learning for distributed load forecasting: Addressing data imbalance in smart grids," in *The Fifteenth International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies*, IARIA, 2025, pp. 1–2.

[2] A. Jain, H. Patel, L. Nagalapatti, N. Gupta, S. Mehta, S. Guttula, et al., "Overview and importance of data quality for machine learning tasks," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 3561–3562.

[3] H. Habbak, M. Mahmoud, K. Metwally, M. M. Fouda, and M. I. Ibrahem, "Load forecasting techniques and their applications in smart grids," *Energies*, vol. 16, no. 3, p. 1480, 2023.

[4] GDPR, "General data protection regulation," *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC*, 2016.

[5] P. L. Ambassa, A. V. Kayem, S. D. Wolthusen, and C. Meinel, "Inferring private user behaviour based on information leakage," *Smart Micro-Grid Systems Security and Privacy*, pp. 145–159, 2018.

[6] G. Wood and M. Newborough, "Dynamic energy-consumption indicators for domestic appliances: Environment, behaviour and design," *Energy & Buildings*, vol. 35, pp. 821–841, 2003.

[7] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet, and D. Irwin, "Private memoirs of a smart meter," in *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*, 2010, pp. 61–66.

[8] J. Chen, H. Yan, Z. Liu, M. Zhang, H. Xiong, and S. Yu, "When federated learning meets privacy-preserving computation," *ACM Comput. Surv.*, vol. 56, no. 12, Oct. 2024, ISSN: 0360-0300. DOI: 10.1145/3679013.

[9] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 54, PMLR, Apr. 2017, pp. 1273–1282.

[10] L. Li, Y. Fan, M. Tse, and K.-Y. Lin, "A review of applications in federated learning," *Computers & Industrial Engineering*, vol. 149, p. 106 854, 2020.

[11] P. M. Mammen, "Federated learning: Opportunities and challenges," *arXiv preprint arXiv:2101.05428*, 2021.

[12] J. Wen, Z. Zhang, Y. Lan, Z. Cui, J. Cai, and W. Zhang, "A survey on federated learning: Challenges and applications," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 2, pp. 513–535, 2023.

[13] J. Jithish, B. Alangot, N. Mahalingam, and K. S. Yeo, "Distributed anomaly detection in smart grids: A federated learning-based approach," *IEEE Access*, vol. 11, pp. 7157–7179, 2023.

[14] H. Liu, X. Zhang, X. Shen, and H. Sun, "A federated learning framework for smart grids: Securing power traces in collaborative learning," *arXiv preprint arXiv:2103.11870*, 2021.

[15] X. Cheng, C. Li, and X. Liu, "A review of federated learning in energy systems," *2022 IEEE/IAS industrial and Commercial Power System Asia (I&CPS Asia)*, pp. 2089–2095, 2022.

[16] A. GroSS, A. Lenders, F. Schwenker, D. A. Braun, and D. Fischer, "Comparison of short-term electrical load forecasting methods for different building types," *Energy Informatics*, vol. 4, no. S3, Sep. 2021, ISSN: 2520-8942. DOI: 10.1186/s42162-021-00172-6.

[17] A. Fayyazbakhsh, T. Kienberger, and J. Vopava-Wrienz, "Comparative analysis of load profile forecasting: Lstm, svr, and ensemble approaches for singular and cumulative load categories," *Smart Cities*, vol. 8, no. 2, p. 65, 2025.

[18] A. M. N. Ribeiro, P. R. X. do Carmo, P. T. Endo, P. Rosati, and T. Lynn, "Short-and very short-term firm-level load forecasting for warehouses: A comparison of machine learning and deep learning models," *Energies*, vol. 15, no. 3, p. 750, 2022.

[19] K. Ullah, M. Ahsan, S. M. Hasanat, M. Haris, H. Yousaf, S. F. Raza, et al., "Short-term load forecasting: A comprehensive review and simulation study with cnn-lstm hybrids approach," *IEEE Access*, 2024.

[20] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konený, et al., *Adaptive federated optimization*, 2021. arXiv: 2003.00295 [cs.LG].

[21] X. Wu, F. Huang, Z. Hu, and H. Huang, *Faster adaptive federated learning*, 2023. arXiv: 2212.00974 [cs.LG].

[22] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, 2020.

[23] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning*, PMLR, 2020, pp. 5132–5143.

[24] C. Ma, J. Li, M. Ding, H. H. Yang, F. Shu, T. Q. Quek, et al., "On safeguarding privacy and security in the framework of federated learning," *IEEE Network*, vol. 34, no. 4, pp. 242–248, 2020.

[25] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," *Future Generation Computer Systems*, vol. 115, pp. 619–640, 2021.

[26] M. N. Fekri, K. Grolinger, and S. Mir, "Distributed load forecasting using smart meter data: Federated learning with recurrent neural networks," *International Journal of Electrical Power & Energy Systems*, vol. 137, p. 107 669, 2022. DOI: https://doi.org/10.1016/j.ijepes.2021.107669.

[27] A. Taïk and S. Cherkaoui, "Electrical load forecasting using edge computing and federated learning," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, IEEE, 2020, pp. 1–6.

[28] Y. Liu, Z. Dong, B. Liu, Y. Xu, and Z. Ding, "Fedforecast: A federated learning framework for short-term probabilistic individual load forecasting in smart grid," *International Journal of Electrical Power & Energy Systems*, vol. 152, p. 109 172, 2023.

[29] Y. Shi and X. Xu, "Deep federated adaptation: An adaptative residential load forecasting approach with federated learning," *Sensors*, vol. 22, no. 9, p. 3264, 2022.

[30] C. Briggs, Z. Fan, and P. Andras, "Federated learning for short-term residential load forecasting," *IEEE Open Access Journal of Power and Energy*, vol. 9, pp. 573–583, 2022.

[31] R. Rahman, N. Kumar, and D. C. Nguyen, "Electrical load forecasting in smart grid: A personalized federated learning approach," in *2025 IEEE 22nd Consumer Communications & Networking Conference (CCNC)*, IEEE, 2025, pp. 1–2.

[32] *Smart meters in london*, https://www.kaggle.com/datasets/jeanmidev/smart-meters-in-london, Accessed: 2010-09-30.

[33] A. Wallis, U. Ludolfinger, S. Hauke, and M. Martens, "Towards federated short-term load forecasting," *Internationale Energiewirtschaftstagung (IEWT 2021)*, pp. 1–9, 2021.

DRONE-SLAM: Dense Registration and Odometry for Near-field Environments with UAVs

Diego Navarro [1,2], Ezio Malis [2], Raphael Antoine [1], and Philippe Martinet [2]
[1]ENDSUM team. Cerema Normandie-Centre, France
[2]ACENTAURI team. Centre Inria d'Université Côte d'Azur, France
e-mail: diego-navarro.tellez@inria.fr

***Abstract*—Structural health monitoring applications (SHM) require large volumes of data that can only be acquired with automatic methods, such as robots or UAVs. However, localizing a drone in close proximity to the target structure poses significant challenges, particularly when the GPS is not precise enough. A possible solution is to use vision-based localization in a previously build model of the target. The localization can be however very challenging if, for safety reasons, the reconstruction of the model has been acquired far from the target and the target has a low-texture. These issues often result in reduced data density and obstruct convergence to the true position. The majority of previous work deal with the localization at far-field regions, while only few methods address the vision-based localization close to targets with low-texture. This paper presents a novel Dense Visual SLAM method tailored for close-range localization to surfaces using unmanned aerial vehicles (UAVs) in GPS degraded conditions. Our method uses a custom registration method to enable realistic rendering with dense maps, designed for close-range visual odometry and surface modeling. The system operates in two steps: First, the UAV performs an exploratory flight with a stereo camera to build a dense map, modeling surfaces as ellipsoids. Secondly, the system exploits the map to generate reference data, enabling dense visual odometry (DVO) in close proximity to the surfaces without the need of stereo data. Experiments in realistic simulated environments demonstrate the system's capability to localize the drone within 16 cm accuracy at a distance of 2 m from the surface, outperforming existing state-of-the-art approaches. Tests on real data confirm this performance in real low-texture scenarios.**

Keywords-robotics; autonomous vehicles; vision and scene understanding; volumetric image representation.

I. Introduction

This work builds upon the hybrid SLAM method introduced in [1], keeping the two-step workflow while incorporating significant advancements. Such methodological improvements reflect the adaptation of localization techniques for UAV-based monitoring and inspection tasks. UAVs have experienced a rapid increase in adoption across diverse fields [2] in recent years. Natural hazards and civil engineering sectors are no exception, with growing enthusiasm for utilizing these advanced tools in classification surveys and inspections [3][4]. UAVs can operate either manually, with pilots directly controlling the drone, or autonomously, using technologies such as environmental sensing and self-localization. Recent technological advancements have led to the development of autonomous drones capable of navigating intricate and geometrically complex environments [5]. As a result, their role has evolved in applications such as infrastructure inspections, cliff surveys, and more generally in environmental monitoring [6].

While some surveilling applications can be performed with current UAV localization techniques [7], more demanding tasks like close-range inspection are still an issue. Such tasks enable a more effective analysis and diagnosis through SHM techniques [8]. With richer data, like radar measurements ([9], [10]), the quality of structure's digital twins and their pertinence can be greatly improved. However, these measurements require proximity to the surface, which is often done manually.

Manual acquisition not only poses significant risks to human agents but also lacks in coverage. Although UAVs have the potential to solve this problem but there are some challenges to address first. Accurate localization is a critical requirement for deploying autonomous robots for SHM tasks. However, conventional techniques often struggle with the accuracy of the pose estimation, specially in cluttered environments or close to structures.

Traditionally, UAVs are located using GPS or even GPS-RTK modules, while they are convenient, their performance depends heavily on the operation conditions. There are two principal issues. First, the coverage of the GPS signal, which can deteriorate or even fail sometimes. In the case of GPS RTK, the effects of bad reception can degenerate by the proximity to the correction transmission station and the availability of correction data. Secondly, obstacles between the signal sources and the reception antenna can greatly impoverish the quality of the location. Multiple obstacles, typically in urban environments, produce multi-path interference, further degrading localization quality [11].

When GPS signals are unavailable or unreliable, V-SLAM provides a robust alternative for UAV localization. Visual data can serve for localization without worrying about multi-path interference or data availability. However, there are challenges to solve before being able to locate an UAV during close-range inspections. UAVs face strict weight and energy constraints, which limits their capacity to carry both measurement and high-performance localization equipment simultaneously. This can be mitigated by dividing the computational load in two steps: A mapping step with a high-capacity computation payload and a second step with a simplified tracking unit and measurement equipment. With the deactivation of the mapping module during the second step, the tracker can only rely on a pre-generated map as reference.

The lack of map updates introduces other challenges, localizing far from the original map trajectory often results in reduced tracking accuracy. The reasons may vary depending

on the method, one of the most common for feature-based V-SLAM methods is the perspective narrowing. As cameras approach the inspection surface, their field of view narrows, reducing available visual information and rendering previously mapped features unrecognizable from the new perspective.

In this paper, we present a V-SLAM framework tailored for close structure inspection using UAVs. To address the challenges of this use case, we propose a two-step V-SLAM framework. First, a stereo camera captures images during a mapping flight, using a feature-based method for ego-pose estimation. The system's modular design allows flexibility on the choice for mapping ego-localization methods.

The resulting dense map serves dual purposes: aiding mission planning and enhancing localization in the second step. In this phase, precise localization is performed using a lightweight monocular camera, reducing the weight and energy demands typically associated with additional measurement equipment like radar or thermal cameras. Key contributions of our method are:

- A registration method that enhances accuracy and robustness of the localization system.
- Generation of dense map that enables multi-session localization, increasing system versatility.
- Use of dense map for precise localization in close-range inspection scenarios.
- A EWA volume splatting variant, tailored for low density point-cloud rendering.
- Support for agent localization across mixed hardware configurations.

The remainder of this paper is organized as follows: Section II reviews related work that influenced this study. Section III describes the proposed method, with details of the mapping workflow presented separately from the localization modules. Section IV presents the experimental results that validate system performance, along with the test conditions. Finally, the conclusion discusses the results and outlines future research directions to improve the method.

II. Related work

While some existing methods attempt to handle issues like perspective narrowing, few specifically address the challenges posed by scale changes under extreme conditions. Classic V-SLAM approaches propose different strategies for pose estimation. On one hand there are the feature-based methods, such as ORB-SLAM and its successors [12][13], are widely used due to their computational efficiency. However, they perform poorly in low texture scenarios, specially in localization mode, where the system conserves computational resources by relying on pre-generated maps rather than real-time mapping.

On the other hand, dense direct methods [14], [15], derived from SfM techniques (Structure from motion) are capable of locating the agent's location and demonstrate robustness in low-texture regions. This robustness derives form the use of pixel intensity gradient over the entire image to better estimate the pose changes, particularly in conditions where feature-based methods may struggle. However, the reduced field of view near inspection surfaces not only decreases point density, it also introduces gradient inconsistencies, which can compromise the system's localization performance.

Advances in V-SLAM have explored mixed approaches to benefit from the strengths of both feature-based and dense methods. Prior work has explored the creation of dense maps from sparse Key Frames, as presented in [16]. Zhang and Shu integrated a dense mapping component into the ORB-SLAM2[12] framework using stereo data. While effective for merging overlapping point cloud regions, this approach does not address localization challenges in close-range inspections. Specifically, the ORB-SLAM tracker struggles to match descriptors in such environments, likely due to the limitations of scale invariance, where descriptors lose discriminative power at extreme proximity. Moreover, the map generated by Zhang et al.'s dense mapping thread neither addresses this problem nor demonstrates the necessary capabilities to resolve it.

One approach to mitigate perspective changes is to improve the ability to generate pose hypotheses from multiple viewpoints. For example, Kerb et al. [17] proposed a novel modeling method that enables realistic rendering by optimizing 3D ellipsoid parameters using outputs from SfM algorithms. These ellipsoids can be rendered from new viewpoints via Gaussian splatting, an old rasterization technique that is now more affordable thanks to advancements in hardware architectures and rendering methods. This work reignited interest in dense mapping approaches, leading to the emergence of new SLAM techniques inspired by [17]. With proper equipment, VSLAM methods can benefit from more realistic images, increasing their robustness in far-field regions of the map.

Although Gaussian-splatting can mitigate the perspective narrowing issues, it does not guarantee the geometric accuracy of the ellipsoids because its optimization focuses on visual quality rather than spatial precision. Moreover, high quality images remain computationally expensive, even with an optimized implementation. While effective in structured and information-rich environments, they are limited in low texture scenarios. Additionally, their energy consumption and hardware requirements make them impractical with UAV platforms.

Compared to the initial paper [1], this paper introduces new improvements focused to improve the robustness of the system in close-range observations. First, the rasterization method is improved, transitioning from an ellipsoid representation based on the statistical properties of the observations to a model that prioritizes geometric accuracy and incorporates occlusion handling. Secondly, this work expands on the graph-based map management introduced in the previous paper by detailing the logic behind node creation and adding a critical feature: the detection and handling of pixel resolution changes to enhance DVO performance This enhancement enables DRONE-SLAM to detect when incoming images improve the surface model's detail and prioritize their integration into the map. Finally, this paper introduces new experimental validation with real-world data acquired in conditions relevant to close-range

UAV inspections. While [1] primarily focuses on generating a dense map to address the point density problem in close-range observations, this paper extends the approach significantly. It emphasizes geometric accuracy, realistic scene rendering, and detailed implementation, bringing the system closer to a fully integrated dense SLAM framework.

III. Proposed method

To address the challenges of close-range structure inspection we propose a visual localization system. In contrast to GPS-localization that can easily be occluded by the elements in the environment, this kind of methods do not rely solely in fixed sources of information and use the environment in their favor. While the system can work with conventional cameras, it was designed for the application case of UAV deployment. In the current state of implementation, the proposed method is designed to work in two steps. This section explains the main elements of the proposed method. First, we will elaborate over the details of the mapping step, the workflow, how the 3D data is computed and the aggregation of the different key-frames. Then, the elements of the scanning step will be explained in the scanning section along with other key elements of the workflow related to the estimation of the drone pose relative to the environment.

A. Mapping

In the first step the pilot of the mission shall perform a simple exploratory fight around the structure at a safe distance with the mapping payload composed by a pair of synchronized stereo cameras. The main components of the mapping workflow (see Figure 1) are described as follows:

1) System initialization: In this implementation, the UAV is supposed to start close to a common takeoff position for both mapping and scanning steps. This assumption is necessary because, currently, the system lacks a loop closure module capable of recognizing previously visited locations. During the first scene observation, the mapping module computes the first data-cloud from the stereo images and saves it as the first reference data. The selection of the initial observation is crucial for high-quality mapping, as the rest of the map will be forced to be coherent to the scale of this frame. This initial data-cloud is assumed to provide a fair reference, as the flight plan includes a preliminary parameter verification. This can either be done with a calibration check or an on-site adjustment of the depth estimation module's parameters.

2) Depth estimation: The current implementation relies in two different methods for scene depth estimation. For low-texture unstructured scenarios, the depth map is computed using the stereo Block Matching (BM) algorithm provided by the OpenCV library [18]. With proper parametrization, this method can approximate the overall geometry of the scene even in low texture scenes. However, its performance depends heavily on the parametrization, which has to be done manually for each scenario, compromising mapping precision. Moreover, the disparity estimation quality of this module vary depending on the scenario, requiring additional adjustments. In case of very low-textured environments, this algorithm is obliged to chose between sensibility for geometry detection and noise-free depth-maps. That said, the use of this module is provisional until the implementation of a better method is achieved.

The second module used is the CREStereo ML model, proposed by Li et al. [19], which performs well in structured scenarios, offering a full depth-map estimation (BM often skips parts of the images when the texture uniqueness parameter is too strict). For this implementation we used a pre-trained version of the model available on the project's GitHub repository. However, as this model was not specifically trained for our unstructured environments, often mistakes the surfaces of natural formations (i.e. cliff scenario) by planes.

In both cases, the disparity images are collected and used to compute a depth-map that will then be used to build the data-clouds that from now on we will call Surfaces ($\mathbf{S}_x$).

3) Environment modeling: These data-structures model more than a simple point-cloud. While many point-cloud registration methods assume that cameras, very much like lidars, observe a set of perfect 3D points in the space.

Our Surface model, however, considers the fact that cameras are sensors that discretize the space in pixels. Due to technical restrictions linked to the construction of the camera sensor, pixels cannot be infinitely small. In consequence, this data-structure considers each pixel as a patch of the surface that is observed at a certain position to conserve as much information as possible from the observation model. To implement the patch hypothesis, each j pixel is saved with the following data:

- 3D mean position $\mathbf{m}_j$ computed and updated through map registration
- A shape matrix $\mathbf{M}_j$ representing the patch covered by the pixel
- Color information of the pixel $\mathbf{c}_j$

The shape matrix of each pixel $\mathbf{p}_j$ is computed so the ellipsoid can cover a plane matching the size of the pixel in the real world (see Figure 2):

$$w_{px_j}(\mathbf{p}_j) = \frac{d_j w_{cam}}{w_{img} f_{length}} \tag{1}$$

$$h_{px_j}(\mathbf{p}_j) = \frac{d_j h_{cam}}{h_{img} f_{length}} \tag{2}$$

Here, d_j is the depth value at $\mathbf{p}_j$, w_{cam} and h_{cam} are the camera sensor size in meters, w_{img} and h_{img} are the image size in pixels and f_{length} is the focal length in meters. The shape matrix can be composed as follows:

$$\mathbf{M}_j = \begin{bmatrix} \left(\frac{w_{px_j}}{\sqrt{2}}\right)^2 & 0 & 0 \\ 0 & \left(\frac{h_{px_j}}{\sqrt{2}}\right)^2 & 0 \\ 0 & 0 & \left(\frac{h_{px_j}+w_{px_j}}{2}\right)^2 \end{bmatrix} \tag{3}$$

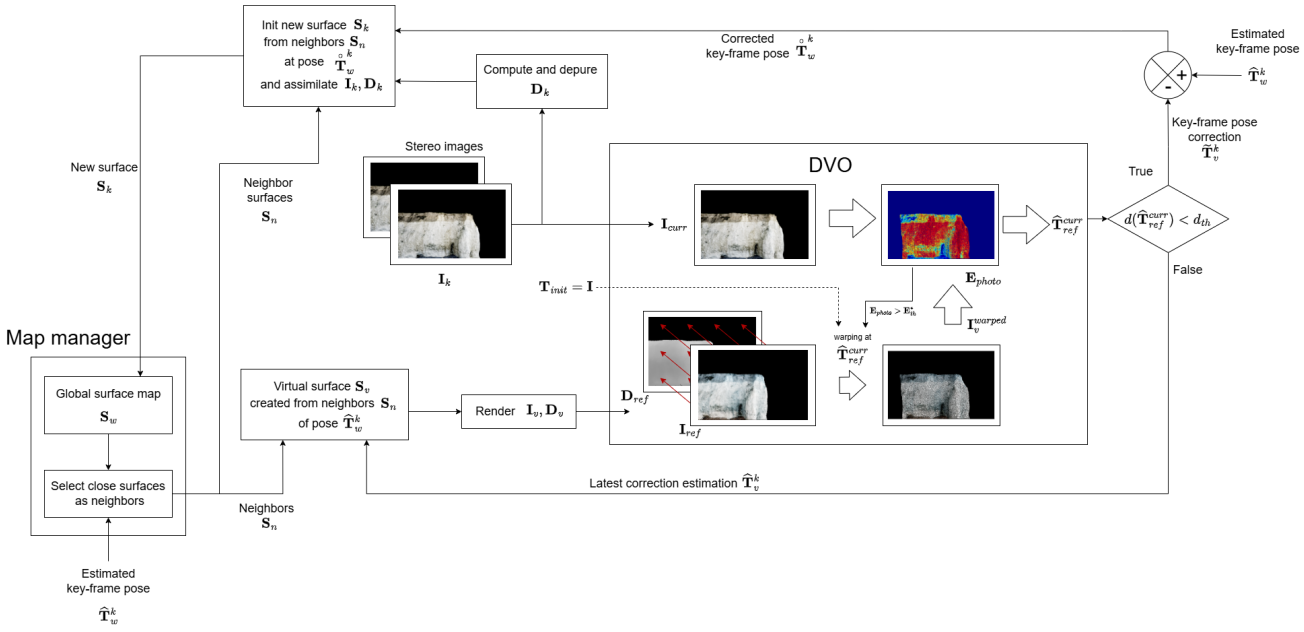


Figure 1. Mapping thread workflow. Input images $\mathbf{I}_k$ are frames selected when the current estimated pose $\widehat{\mathbf{T}}_w^n$ checks any criteria for node creation becoming then $\widehat{\mathbf{T}}_w^k$. DVO module is initialized with identity as the estimated pose and the stereo-images pose should be the same. Tracking thread uses similar structure but instead using virtual reference data $\mathbf{I}_v, \mathbf{D}_v$ the DVO module uses the data stored in the closest surface to the current estimated pose $\widehat{\mathbf{T}}_w^n$.

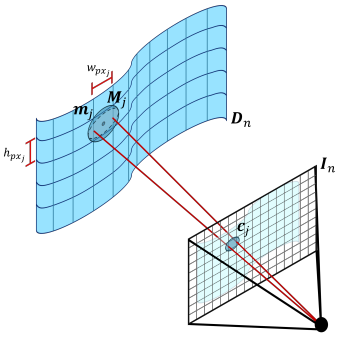


Figure 2. For each n incoming frame there is a color image $\mathbf{I}_n$ and its respective depth-map $\mathbf{D}_n$, the shape matrix is computed so the corners of the pixel plane are tangent to the ellipse border. The surplus ellipsoid coverage is assumed to be forget by the weighted average.

The shape matrices are first computed in the camera frame and then rotated to the global frame so they can be updated with the next observations. The 3D position of the ellipsoid is determined by the weighted average of the point observations along the camera axis. The color information is saved in the form of CIELAB L,A,B coefficients and also fused trough weighted average. This approach contemplates homogeneous lighting across the scenes, further work on light modeling will improve the robustness of the system in more realistic lighting conditions.

4) Pose correction: As the agent moves, new stereo images are continuously acquired. These new observations are used by the tracking module to keep an estimation of the pose of the agent. Simultaneously, the mapping module will select specific frames as *key-frames* $\mathbf{I}_k, \mathbf{D}_k$. At each k selected frame, the incoming stereo images are aligned with the closest reference surface in the map. To achieve this, a virtual observation form the actual estimated pose is computed using the rasterization module. The rendering of the virtual observation will be explained in Section III-C.

Once rendered, the dense odometry module (DVO) computes the pose between the virtual observation of the map $\mathbf{I}_v$, generated at the current estimated pose $\widehat{\mathbf{T}}_w^k$, and the current stereo frame $\mathbf{I}_k$. The resulting pose, $\widetilde{\mathbf{T}}_k^v$, represents the error of the tracking module. In consequence, the true pose of $\mathbf{S}_s$ can be computed as:

$$\overset{\circ}{\mathbf{T}}{}_w^k = \widehat{\mathbf{T}}_w^k \left(\widetilde{\mathbf{T}}_v^k\right)^{-1} \tag{4}$$

Since the visibility of the map ellipsoids depends on the position of the camera, the computation of the correction $\widetilde{\mathbf{T}}_v^k$ is done iteratively. This continues until the position and rotation of the pose are under a threshold defined by the user. For each iteration, a new set of virtual reference image $\mathbf{I}_v$ and depth-map $\mathbf{D}_v$ are computed (feedback loop at the bottom of Figure 1).

5) Surface data-structure: Once the pose is corrected, the surface is supposed aligned with the global map. Under this assumption, information in the incoming frame can either be used to update an existing reference surface or to create a new one.

Initialization: New surfaces are initialized with data from the neighbors and the incoming frame that triggered the creation of the new surface. Since all frames are aligned, neighbor data

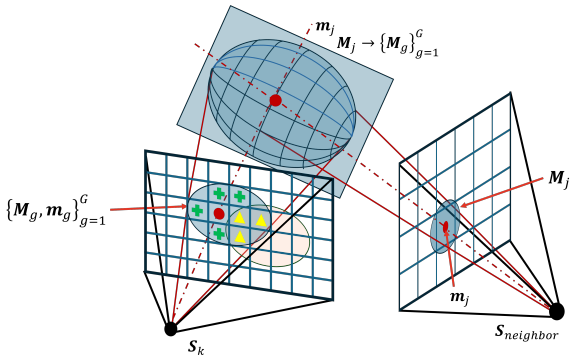


Figure 3. Reference surfaces are initialized with neighbor's data. For instance, the shape matrix $\mathbf{M}_j$ member of a neighbor surface $\mathbf{S}_N$ is projected into $\mathbf{S}_k$ using a linearized projection model around $\mathbf{m}_j$. Perfect correspondences are passed into the grid (red circle), unoccupied pixels covered by the projection are initialized with new ellipsoids (green crosses), occupied cells (orange ellipsoid) are updated following the initialization criteria (yellow triangles). When multiple cells are covered by the projection, the ellipsoid $(\mathbf{M}_j, \mathbf{m}_j)$ becomes the group of the G ellipsoids covered by the projection: $\{\mathbf{M}_j, \mathbf{m}_j\}_{g=1}^{G}$. In $\mathbf{S}_n$ the group will be updated collectively but in $\mathbf{S}_{new}$ each cell preserves individuality.

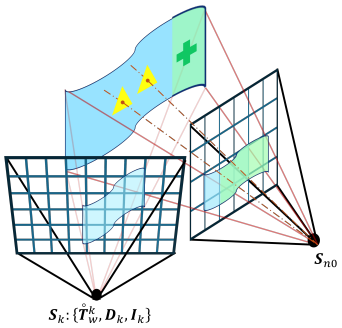


Figure 4. After initialization or when the incoming key-frame does not meet any of the criteria to create a new node, the data is assimilated in the closest reference surface $\mathbf{S}_{n0}$. The selected data $\mathbf{D}_k$ and $\mathbf{I}_k$ are projected into the camera frame of $\mathbf{S}_{n0}$ using an intermediate surface $\mathbf{S}_k$. Unoccupied ellipses are added into the grid (green) and existing ones are updated along the camera axis of $\mathbf{S}_{n0}$ (yellow). We call this operation retro-splatting since a 3D reconstruction of the incoming data in camera frame is *splatted back* into the surface frame.

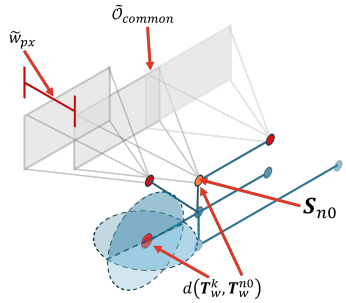


Figure 5. There are different ways to trigger the creation of a new node: The distance inter-surface $d(\mathbf{T}_w^k, \mathbf{T}_w^{n0})$, the pixel size difference $\widetilde{w}_{px}$ and the rate of overlap $\widetilde{\mathcal{O}}_{common}$. These metrics are computed with respect to the closest surface $\mathbf{S}_{n0}$ data.

can be passed to the new surface through a simple projection into the current camera frame (see Figure 3).

In each frame, the data is organized as a grid of ellipsoids to simplify the manipulation of the data, this grid emulates the image plane. As each neighbor pixel is transferred into the new surface, the ellipsoids are projected following the algorithm proposed in [20]. To achieve this, the image projection matrix is linearized at the center of the ellipsoid. Then, based on the dimensions of the ellipsoid, the coverage of cells is computed to determine whether a reference to the neighbor cell is placed or a new ellipsoid is created.

Since ellipsoids might cover more than one pixel, new instances are created and stored in the new gird. In consequence, the old instance is divided to contain the group of ellipsoids generated by the projection. Now $\mathbf{M}_j$ becomes $\mathbf{M}_G = \left\{\mathbf{M}_i^{(g)} | g = 1, \cdots, G\right\}$, as shown in Figure 3. This ensures that updates to the current surface affect each individual ellipsoid, while updates form a point of view that sees the group as one pixel will affect all the ellipsoids at once. When multiple ellipsoids fall in the same pixel, they are also grouped but only if they are close enough. If they are farther than a certain threshold, only the closest one is kept.

This maintains spatial relationship of adjacent pixels since the update of only visible pixels might derive in noisy surfaces.

Update: Once the new surface has all the neighbor ellipsoids that can be projected onto this plane, the surface assimilates the information coming from the current camera observations $\mathbf{I}_k, \mathbf{D}_k$ (see Figure 4). First, the position of the ellipsoids are updated pixel by pixel, if the cell is empty, a new ellipsoid will be created. The update of the position of the ellipsoid is done computing a weighted average of the pixel depth values along the camera axis. This simplify computations and ensure the respect of the projection model and the alignment of the ellipsoids of the map in the image plane. The current implementation updates the conic matrices with a weighted average in order to get a new matrix that can cover both old and new ellipsoids.

Is worth noting that, during the initialization of a new surface the depth-map $\mathbf{D}_k$ and the color image $\mathbf{I}_k$ are assimilated directly since they are observed in the pose that triggered the creation of the new surface. However, in the case of the update of an existing surface, the depth-map is used to create an intermediate surface $\mathbf{S}_k$ that will then be used to render a new set of $\mathbf{I}_{vk}$ and $\mathbf{D}_{vk}$ observed from the reference surface pose. In summary, during the update and initialization an operation of retro-splatting is used to project the new information into the grid, either to update the surface or to initialize it. After full initialization, the new surface $\mathbf{S}_{new}$ becomes member of the map nodes $(\mathbf{S}_w)$.

6) Mapping criteria: One of the main objectives of this algorithm is to be able to retain as much as information as possible to enable the localization at proximity to surfaces. With this in consideration, the system includes a set of criteria to implement this policy (see Figure 1). Each condition is verified at the end of the alignment of a new key-frame. First, the graph will trigger the creation of a new surface when any distance from the last reference surface $(d_t(\mathbf{T}_w^k, \mathbf{T}_w^{n0})$ or

$d_R(\mathbf{T}_w^k, \mathbf{T}_w^{n0})$) exceeds a threshold configured by the user or when no reference surface is found:

$$d_t(\mathbf{T}_w^k, \mathbf{T}_w^{n0}) = \|\mathbf{W}_{\text{pos}} \cdot (\mathbf{t}_k - \mathbf{t}_{n0})\| \quad (5)$$

$$d_R(\mathbf{T}_w^k, \mathbf{T}_w^{n0}) = 2 \cdot \arccos(\mathbf{q}_k \cdot \mathbf{q}_{n0}) \quad (6)$$

Where $\mathbf{t}$ is the position component, $\mathbf{q}$ the orientation in the form of unitary quaternion and $\mathbf{W}_{\text{pos}}$ is a set of weights to adjust axis sensibility if needed. This metric is designed to ensure tracking continuity and decrease the accumulation of errors product of the projection pixel-discretization.

Secondly, there is the overlap criterion $\widetilde{\mathcal{O}}_{common}$:

$$\widetilde{\mathcal{O}}_{common} = \frac{|\mathcal{O}_{n0} \cap \mathcal{O}_k|}{|\mathcal{O}_{n0}|} \geq \tau \quad (7)$$

Here, $\mathcal{O}_k$ represents the occupied pixels in the grid of $\mathbf{S}_k$ and τ is an user defined threshold. This policy exits to ensure continuity of the mapping operation, specially during the first stages of the mission. Since there is few information stored in the map, the lack of overlap can cause the DVO module to fail, even if the distance traveled is not long. With this criterion, the graph is prone to detect when new zones are discovered to ensure sufficient conditions for the DVO module to work. At the same time the distance threshold can be set higher to reduce unnecessary node creation.

Then there is the resolution criterion $\widetilde{w}_{px}$:

$$\widetilde{w}_{px} = \frac{\frac{1}{J_{n0}} \sum_{j=1}^{w_{\text{img}} h_{\text{img}}} w_{px_j}^{(n0)}}{\frac{1}{J_k} \sum_{j=1}^{w_{\text{img}} h_{\text{img}}} w_{px_j}^{(k)}} \quad (8)$$

Here, $w_{px}^{(k)}$ represents the size of the pixel in milliliters when projected in the real world (see Figure 2) and J_{n0} represents the total count of pixels with valid depth at the surface $\mathbf{S}_{n0}$. As the agent evolves in the environment, existing data can often be observed in better conditions. Since the objective of the map is to enable precise localization at close-range, the graph manager prioritizes the creation of new nodes when the difference of resolution exceeds a threshold defined by the user. In the current implementation, a resolution improvement of 50% over the original resolution triggers the creation of a new node. The computation is similar to the computation of the ellipsoid shape. The mean pixel size of the incoming frame is compared to the mean pixel size of the closest reference surface as long as the overlap between them is greater than a user defined threshold. This ensures that the new node genuinely represents an improvement in resolution and overrides the distance criterion when necessary. In this special case, ellipsoids are overwritten with new information mixed from previous observations but prioritizing the storage of the new high resolution data to redefine existing data.

B. Performance

The current implementation of the mapping thread processes each frame in 355 seconds, with most of this time spent on creating and manipulating point instances. While most operations take few milliseconds, the creation of point instances and the application of geometric transformations alone account for 181 seconds of the total processing time. The remaining execution time is distributed between image rasterization (11 seconds) and the odometry implementation (30 seconds), both of which are slowed by the creation of intermediate surface structures. Odometry time includes not only the operations of the DVO module but also pose estimation iterations, virtual surface creation. Each iteration also requires a rasterization call, further contributing to the processing time. While similar operations on arrays of data points can be up to 30 times faster, the need to maintain consistent point references across different surfaces limits the use of certain Python optimization tools.

As SLAM systems require an execution time close to the real time for deployment in robotic applications, this algorithm still needs to improve its execution time to be able to be deployed in a real drone. However, this execution time can be greatly improved when implemented in proper conditions. The current version has been written in Python 3, data manipulation is done with the numpy library accelerated with the help of the numba library in some operations.

A significant limitation of the prototype comes from implementation language. The first issue is that Python usually executes a single tread process which is not adequate for the application. While threading is possible at big scale (i. e., running tracking and mapping in parallel), many potentially parallelizable tasks remain executed sequentially. The majority of the operations are constrained to a single thread due to numba's compatibility only with native numpy objects and a limited subset of operators. Furthermore, at the beginning of the call, the interpreter still has to execute the numba overhead to use the compiled executable of the code. Although this overhead is shorter than running un-optimized Python code, it remains slower than a fully multi-threaded implementation in a compiled language.

Another limitation derived from the choice of the programming language is the memory management. To avoid data redundancy, each ellipsoid is modeled as an object instance that contains all the information to describe the geometric entity. When creating a surface, hundreds of thousands of instances have to be created at each node and this operation consumes more time and memory than equivalent operations in C or C++. For instance, with a resolution of 1280x720 pixels, the map manager takes approximately 60 seconds to create a surface and initialize it with the information of its neighbors (dependent on the number of neighbors). This process cannot be parallelized with numba since this library does not work with Python object instances. By contrast, implementing this operation in C or C++ would allow faster indexing by memory allocation and faster data processing with parallelized processing.

Finally, the implementation of the current version does not use any graphic computation resources. This means that operations like the render of the depth and color images could be greatly accelerated for both tracking and mapping processes. This omission contributes further to the system's suboptimal

performance in its current form. However, with targeted optimizations in language, threading, memory management, and GPU utilization, the proposed algorithm shows strong promise for real-time deployment in robotic systems.

C. Localization

In the next step, the drone is assumed to use the information of the dense map generated from the first flight and perform a scanning flight with the measurement payload. This payload includes any specialized sensing system required for the inspection and a localization camera, which does not need to be stereo.

1) Dense visual odometry (DVO): The alignment of the key-frames and the precise tracking of the system depends on the DVO module [21], provided by the OpenRox library. This module uses a reference image and a depth-map to compute the pose between the reference data and another input image. To integrate this module into the current implementation, a Cython wrapper is used to call its functions, as the module itself is developed in C.

In broad terms, the module minimizes the photometric error function between the current image and the warped reference image until it reaches a threshold. The module uses the zero mean cross-correlation index (ZNCC) to evaluate the quality of the alignment between the warped image and the input image. That said, both tracking and mapping threads reject the estimation if this index falls bellow the 70%. Such a low correlation is interpreted as insufficient similarity between the images, indicating that the estimation may be unreliable for accurate localization.

As mentioned earlier, the module is accessed trough a Python warper of the OpenROX library, thus, no splatting technique during the warping operation. Considering that the warping operation often

2) Pixel splatting: To feed the DVO module either for tracking or for key-frame alignment, the map must be rendered into a virtual image plane. In both cases, the map is projected into the image plane of a virtual camera placed at the last estimated pose using the method of Elliptical Weighted Average (EWA) described in [20]. With this method each j point in the point-cloud has to be associated with a shape matrix representing an ellipsoid $\mathbf{M}_j$. First, the camera projection model is linearized around the 3D coordinates of the ellipsoid center $\mathbf{u}_j$. As indicated in the work of Zwicker et al., the linearized projection model of a pinhole camera around a 3D point $\mathbf{u}_j$ can be obtained with the following Jacobian matrix:

$$\mathbf{J}_j = \begin{pmatrix} \frac{f_{px}}{(\mathbf{u}_j)_z} & 0 & -\frac{f_{px}(\mathbf{u}_j)_x}{(\mathbf{u}_j)_z^2} \\ 0 & \frac{f_{px}}{(\mathbf{u}_j)_z} & -\frac{f_{px}(\mathbf{u}_j)_y}{(\mathbf{u}_j)_z^2} \\ \frac{(\mathbf{u}_j)_x}{|\mathbf{u}_j|} & \frac{(\mathbf{u}_j)_y}{|\mathbf{u}_j|} & \frac{(\mathbf{u}_j)_z}{|\mathbf{u}_j|} \end{pmatrix} \tag{9}$$

where f_{px} is the focal distance of the camera in pixels and $(\mathbf{u}_j)_z$ is the z component of the 3D point $\mathbf{u}_j$. This version has been modified from the original text to include the camera intrinsic parameters. Using this matrix, the projected 3D shape matrix $\mathbf{V}_{3D_k}$ can be obtained with the following equation:

$$\mathbf{V}_{3D_k} = \mathbf{J}_k \mathbf{R} \mathbf{M}_k \mathbf{R}^T \mathbf{J}_k^T \tag{10}$$

Here, $\mathbf{R}$ is the rotation matrix in the viewing transformation, which brings the points into the camera frame. A 2×2 subset of the transformed matrix is used to represent the projected 2D ellipse in the image plane. The resulting matrix can now be used to obtain the size of the ellipsoid in the image plane thus the size of the splatting kernel.

The splatting operation is performed at a zone defined by the size of the kernel and the aggregation of color/depth is done using the radial distance metric to ponder the addition. For a point in the kernel $\mathbf{x}_i$, which lies in the kernel of $\mathbf{M}_j$, the radial distance from the ellipsoid center $\mathbf{x}_0$, can be expressed as:

$$r_i(\widetilde{\mathbf{x}}) = \widetilde{\mathbf{x}}_i^T \mathbf{Q} \widetilde{\mathbf{x}}_i \quad \text{where} \quad \widetilde{\mathbf{x}}_i = \mathbf{x}_0 - \mathbf{x}_i \tag{11}$$

Here, $\mathbf{Q}_j$ is the conic matrix, inverse of the shape matrix $\mathbf{M}_j$, which defines the geometry of the ellipsoid in the image plane. To save some computation time, the weights for the aggregation are precomputed and stored into a lookup table w_{tab}. To align the radial distance to the indexes in the lookup table, $\mathbf{Q}_j$ is scaled to match the size of w_{tab}. If $r(\widetilde{\mathbf{x}})$ is bigger than 1 this means that the point is outside the ellipsoid thus does not have to be considered.

In the original code proposed by Zwicker et al. this weights come from the Gaussian distribution expression. However, in this paper it was changed for an exponential decay so the overlap between two close kernels wont generate any blur, particularly at the centers of ellipsoids.

Pixel occlusion When a kernel pixel falls inside the ellipse, the corresponding weight in w_{tab} to determine the weight of the color/depth value in the aggregation. That said, additional considerations have to be observed given the particular case of close-range inspections. Due to the proximity to the surface and the geometry of the scene, ellipsoids that should be occluded may incorrectly change the image if they fall in unoccupied areas of the grid. To deal with the occlusion of ellipses two new weights are added to the traditional weighted average, the occluded term (o_{ded}) and the occluding term (o_{ing}).

$$o_{ded} = \min\left(1, \exp\left(-\frac{1}{2}\frac{d_{old} - d_{new}}{\mathbf{V}_{zz}}\right)\right) \tag{12}$$

$$o_{ing} = \min\left(1, \exp\left(-\frac{1}{2}\frac{d_{new} - d_{old}}{\mathbf{V}_{zz}}\right)\right) \tag{13}$$

Where $\mathbf{V}_{zz}$ is the z component in the shape matrix of the ellipsoid and indicates the sensibility of the decay for this coefficient, the bigger the ellipsoid, slower the occluded value will fade. Inversely, the occluding term will gain in weight as the new point places in front of the old one. Both coefficients are limited to 1 to avoid the saturation of the values as the only interest of these terms is to ponder information weights in the case of occlusion. Once these terms are computed, the

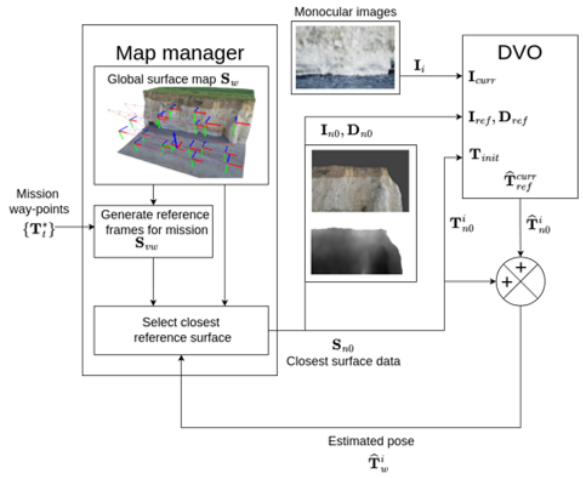


Figure 6. After the mapping flight, a collection of reference surfaces $\mathbf{S}_w$ will be available. Then, according to the trajectory of the flight, a set of specific frames $\mathbf{S}_{vw}$ is also rendered to save computation time around expected poses.

update of the j pixels in the image can be done by an iterative weighted average:

$$\mathbf{c}_j = \frac{o_{ing} w_i \mathbf{c}_i + o_{ded} \mathbf{c}_j}{o_{ing} w_i + o_{ded} w_j} \quad (14)$$

$$\text{where } w_{vj} = w_{tab}(r_i(\widetilde{\mathbf{x}}_i))) \quad (15)$$

Here, pixel j of the Image grid is updated with the values of the kernel pixel i whose weight is defined by its distance to the center of the ellipse. This approach ensures that both occluded and occluding points are accounted for appropriately, preserving the spatial accuracy of the rendered image. A similar aggregation function is applied for the computation of the depth image.

3) Scan workflow: After the mapping flight, the collection of reference surfaces can be exploited for subsequent flights even with reduced optical capacities (monocular instead stereo camera). In *free mode*, each reference surface is rendered at its position to save computation time. Alternatively, if the user decides that the drone has to follow a particular trajectory, a set of mission way-points $\{\mathbf{T}^{t*}_w | t = 1, \cdots, T\}$ must be fed to the map manager (see Figure 6 at the left). Using these points as reference, the map manager can generate a set of virtual positions that ensure the convergence of the DVO module when the agent moves along the vicinity of the desired trajectory. The current implementation takes the shortest path between each pair of way-points and then samples a set of intermediate positions based on a overlap criterion. To do this, the virtual positions are at a user-defined certain distance behind the trajectory to ensure visibility. Their orientation is assumed to be perpendicular to the target trajectory.

The overlap is computed by projecting the path between two way-points and selecting the segment of the path within in the borders of the virtual camera. The next camera center is selected in function of the portion of the projected line covered by the virtual camera and the overlap criterion. The camera FOV overlap is computed assuming the next virtual frame will cover the same portion of the line. This collection of virtual frame poses is also included in the global map as a simplified key-frame, composed only by the color and depth information $\mathbf{I}_{vk}$ and $\mathbf{D}_{vk}$.

Figure 7. Example of the images seen by the simulated drone during the scan flight. As shown in this figure, there are little to no strong details that can help to formulate a correct pose estimation.

During the flight, the map manager continuously selects the closest reference data, key-frame or surface and performs dense odometry relative to the reference key-frame $\mathbf{S}_{vk}$. The DVO module is now initialized with the estimated pose between the reference frame and the last pose estimation $\widehat{\mathbf{T}}^i_w$. If the control data of the agent is available, a preparatory dead-reckoning (DR) step is performed to decrease convergence time (performed in the feedback loop in Figure 6). The current actual implementation the DR step is only for initialization purposes, its implementation into a filtered estimator, such as a Kalman or Particle filter, could provide smoother pose estimations and correctly model the its uncertainty.

IV. Results

A. Simulation data

The performance of the system was first tested in simulations with realistic data collected at the Sainte-Marguerite-sur-Mer cliff (Normandy), monitored in the framework of the Defhy3geo project [22]. The scene is composed of a segment of a cliff model reconstructed using the Agisoft metashape software. The simulation environment was built from the 3D model of the cliff generated form aerial geo-referenced images. A mapping flight was carried on the field, with geo-referenced targets placed to ensure a model alignment accuracy at centimeter-level. The visible area covers a 60x20 meter section of the cliff with a non-structured texture (see Figure 7), providing a challenging environment for texture-based localization.

This test demonstrates the importance of map consistency, particularly its role in maintaining reliable localization results. If the mapping module fails to capture environmental details accurately, an inaccurate initial pose estimate can cause the DVO module to deviate significantly from the correct solution. Despite measures to handle local minima, the lack of sharp geometric features slows pose convergence, potentially reducing precision over time. However, the use of a dense

Figure 8. The video used to test the system presents challenges on multiple levels. First, the lack of significant gradient variations tests the DVO module's performance under extreme conditions. Additionally, the geometry of the environment makes difficult to locate based only in the shape of the surface.

map and realistic rendering in DRONE-SLAM ensures stable performance even under these challenging conditions.

B. Real data

The algorithm was evaluated using a set of real-world data to assess its performance under practical conditions. The test scenario covers a low-texture facade of a building at the Cerema Normandie-Centre institute, providing a challenging environment for visual tracking. Since the drone could not be equipped with a wide-baseline stereo camera, the map was generated using a photometric model built with aerial data from a drone Mavic 3E. This procedure replaces the key-frame selection step of the mapping workflow, providing an opportunity to evaluate the system's ability to exploit other sources of information as reference. Drone operators frequently generate 3D models using similar tools for purposes such as visualization, inspection, or environmental modeling, consistent with the methods applied in this experiment. This test demonstrates DRONE-SLAM's ability to effectively utilize pre-generated models, enabling rapid deployment even on drones with limited payload capacity.

The real dataset was not evaluated with ORB-SLAM as the data collected was not sufficient for a fair comparison. The Mavic 3E cannot be equipped with a wide-baseline camera essential for long-range depth estimation during mapping,. In consequence, the map generated in mapping mode was insufficient to build a significant map, specially when the interest regions lack of details.

As outlined in the proposed workflow, the scanning trajectory and reference data were generated to provide inputs to the tracking thread for pose estimation during the scanning mission. The tracking test used a video recorded by the drone's default wide-lens monocular camera, flying at a distance of 2.5 meters from the wall. The test surface was a plain white wall recorded on a cloudy day (see Figure 8).

The software environment is composed of the following elements:

- Airsim as the simulation environment [23]
- ROS for communication between different software components.
- A Python-based offline registration node for map generation.
- A direct odometry algorithm [21] implemented the OpenROX library (developed by the ACENTAURI team).
- A mission control module to load the map, generate the virtual frames and control the drone.

C. Benchmark conditions

As done in [1], ORB-SLAM3[13] is refereed as comparison method. The principal argument to compare with this method is that this system is relatively popular in robotics since is fast and precise enough for mobile robots deployment. For each scenario, we perform almost the same operation with both ORB3 and DRONE. First a mapping flight is performed with a stereo camera in full SLAM mode, far from the structure. A second flight is performed using the generated map from the first session. In this step ORB3 is configured to run in localization mode, which implies that no new information will be added to the map. That said, is worth noting that the scanning flight with ORB3 is performed still with the stereo camera since with the monocular camera the system struggles to converge. For comparison we take the closest flight to the surface where ORB3 is able to perform the whole flight. The test allows ORB3 to get lost but stops when the error gets above 5 meters without recovery.

D. Experimental results

As presented in [1], the system's performance is evaluated across three aspects: mapping precision, rendered image quality, and localization accuracy. This organization is retained to clearly illustrate the improvements introduced by this method

1) Mapping precision: The first aspect is computed as the distance of the center of the ellipsoids in the map to the reference model. For this we computed a KD-tree to obtain the distances of each surface to the model. To avoid an unfair comparison, points in the ground plane were ignored so the distance represents only the quality of the structure reconstruction. To better appreciate the quality of the map, the distance is expressed in two formats: the quantity of ellipsoid centers that fall into three constraints of precision and the mean distance error of all the points.

TABLE I. MAPPING PRECISION METRICS.

Difficulty	**Lab**	**Cliff**
Hard ($<$ 10 cm)	89.33 %	49.04 %
Medium ($<$ 20 cm)	10.19 %	37.10 %
Light ($<$ 30 cm)	0.46 %	0.13 %
Mean Error Distance	5.17 cm	12.40 cm

2) Rendering: DRONE-SLAM implements improvements to the rasterization method presented in [1], focusing on enhanced rendering accuracy and noise reduction. In [1], a parameter was introduced to define the shape of the weight table, addressing the overlap of Gaussian distributions. In this work, the ellipsoid formulation has been redefined to provide a more accurate representation of the environment. The new implementation uses a splatting method for surface initialization and updates, effectively reducing noise in ellipsoid positioning, enabling a

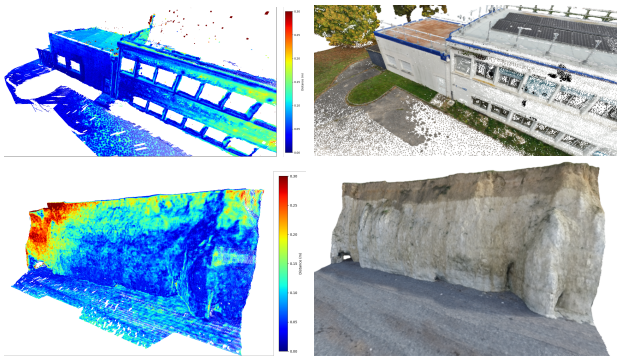


Figure 9. 3D renderings of maps generated by the mapping module (left) with corresponding ground truth models (right). Point colors in the generated maps indicate the distance from the ground truth models. The reference models were generated using geo-referenced images captured during a photometric data-acquisition flight with an average accuracy of approximately 5cm, derived from the GPS-RTK accuracy.

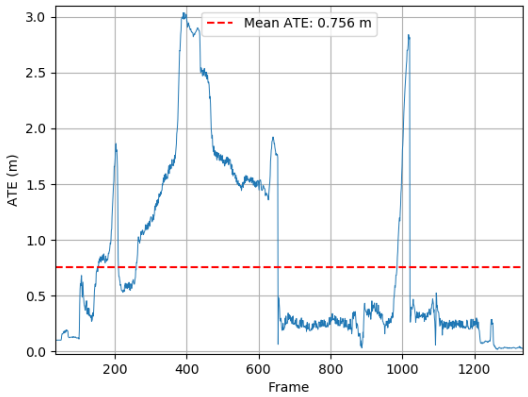


Figure 10. ORB-SLAM3 absolute translation error (ATE) in meters during the scan path with stereo camera. Following the 2 step workflow, ORB-SLAM3 manages to approach up to 5 m close to the cliff surface.

smoother and more consistent data integration. As a result, the mapping method more effectively captures subtle surface details and improves the alignment of data clouds. DRONE-SLAM continues to address the challenges of reduced and uneven point density during close-range observations. However, the need for a custom weight distribution to avoid blurring has been significantly reduced, thanks to improvements in point cloud merging strategies.

3) Localization: simulations: The system's performance is evaluated based on its localization precision during close-range flights to a structure. For this, DRONE-SLAM is compared to ORB-SLAM3 as the baseline in the first simulated scenario. During the scan flight, ORB-SLAM3 initially shows an increase in localization error as the drone approaches the cliff (Figure 10). The system nearly loses tracking capability in the proximity phase. However, as the images contain more informative regions, ORB-SLAM3 recovers. Almost at the end of the flight, a new error peak appears, consequence of a low texture region with oblique geometry that is quickly occluded as the drone follows its path.

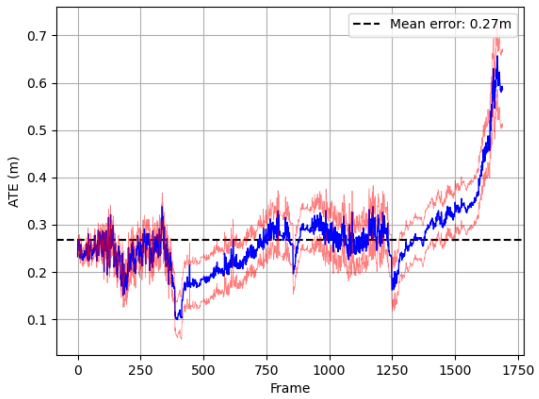


Figure 11. DRONE-SLAM absolute translation error (ATE) in meters during the scan path with monocular camera. The red lines represent the standard deviation cumulated over the time as a from to show the precision of the system. This flight was performed 2 meters away from the cliff.

In contrast, we observe a stable behavior from DRONE-SLAM. The localization error maintains over the different regions of the map. The peak of the end is due to a poor performance of the masked SSIM metric to trigger new nodes when small but relevant regions are discovered. Compared to the performance presented in [1], DRONE SLAM shows more precision since the variations of its estimations remains coherent with the movement of the drone during the scan. Moreover, the peak of error that affected both methods is no longer observed in the case of DRONE (see Figure 11).

4) Localization: real data: Real data tests confirmed that increased resolution, achieved by generating the photometric model at a lower altitude than in the cliff scenario, enhances system performance. In this scenario, the performance comments will be done wrt to the GPS-RTK data synchronized with the video images. As seen at the top of the Figure 12, the pose estimation closely follows the GPS-RTK position estimation. A continuous offset is observed, potentially caused by delays in GPS updates. This observation emerges from the shape of the GPS path that sometimes seems slightly off with respect to the movement observed in the camera. Moreover, the error metric of the odometry seems to increase when these abrupt changes occur. That said, further study on the test conditions has to be done but for the moment, the GPS RTK will be considered as the groundthruth data.

The ATE metric shows a mean error of 17 cm (Figure 13), with stable performance even in low-texture regions. The system performs slightly better during lateral movements compared to vertical movements. As stated before, the lack of geometry features contributes to the ambiguity of the solutions, thus undermining precision. Another factor affecting precision is the use of masks for reference data. Since the reference data consists only of valid map renderings, background elements are often masked and ignored, reducing available visual cues for localization. As a result, the DVO module may observe only planar regions with minimal texture variation, such as

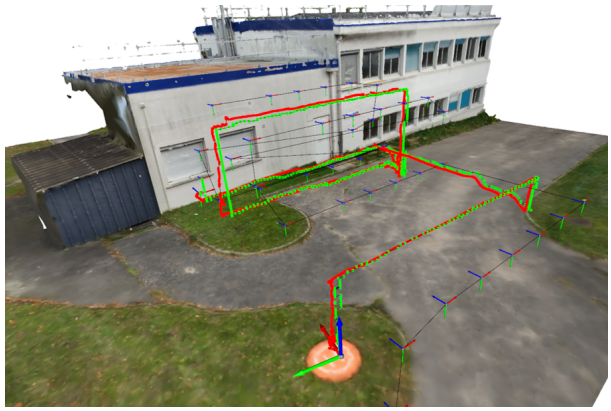


Figure 12. Path followed during the lab flight. GPS-RTK positions are showed in green and DRONE estimations in red. The drone first made a flight at 4 meters from the wall and then another one at around 2m from the wall. The second part of the flight was not completed for KF graph issues.

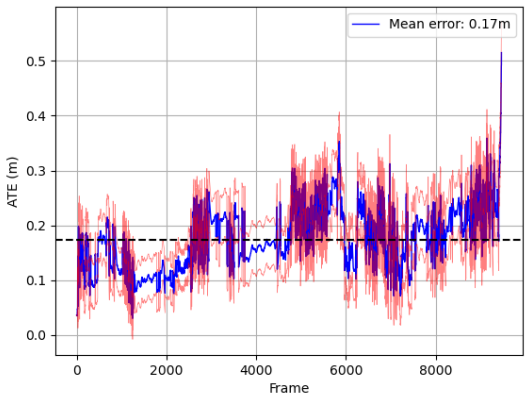


Figure 13. DRONE-SLAM ATE in meters during wall inspection. The system showed better behavior in this structured scenario. The higher performance of the depth estimation module seems to improve the capacity of the mapping module to register subtile details in the texture, which is complicate to perform with relatively noisy estimations like StereoBM.

a uniform wall surface, instead of building edges. These conditions demand high precision to capture subtle texture changes; otherwise, the risk of non-convergence significantly increases.

The system's accuracy heavily depends on the availability of virtual keyframes during the scan step, as these provide essential reference points for pose estimation. This is because the DVO module can diverge if the distance between the reference and the input image are too far. To address this, the system is designed to generate new reference frames when needed, but the mission planner must trigger keyframe generation opportunely. At the end of the sequence, during the flight at 2 meters from the wall, the system failed due to an inadequate keyframe generation strategy in previously unexplored regions. The system performs smoothly in regions where the path planner has pre-generated virtual keyframes, but its performance degrades in unplanned areas. The current implementation relies on inter-keyframe distance criteria to mitigate these issues. However, there are some cases that need a more robust keyframe generation criterion to ensure stability.

V. Conclusion

This paper presented a custom mapping and localization framework, designed for close-range inspection scenarios. The evaluation metrics show that DRONE-SLAM can successfully generate a high resolution dense model of the environment tailored for this use case. Thanks to this dense map and the rendering technique, the system can achieve accurate localization with simple equipment such as monocular cameras. Additionally, the rendering method compensates the perspective narrowing problems while preserving robustness face to low-texture data.

The system's performance was measured across three aspects: First, the quality of the map reconstruction. Secondly, the quality of the custom render method and its compatibility with the DVO module. Finally, the localization accuracy in different scenarios.

While the performance of the system shows an improvement in the quality of estimation in close-range scenarios, some challenges remain. First, the framework must achieve real-time execution to meet the practical requirements of field deployment. Analysis of the execution time profile reveals that the main issues lie in the implementation language and its lack of direct memory access.

Second, the mapping method can improve its robustness in outdoors scenarios. While naive color fusion may work under optimal conditions, the system must reliably handle varying weather. Ongoing work focuses on surface luminance mapping to enable localization with non-uniform lighting. Likewise, this can also improve the multi-session performance of the system as it will allow the use of the map in different conditions and even simulate specific lighting scenarios.

Finally, further research is needed to improve the criteria for generating virtual reference frames. Progress in this aspect will both enhance system stability and extend the utility of dense maps for navigation and mission planning. Simulations show that DRONE-SLAM is more accurate at proximity to the surface with reduced equipment. The test with realistic data demonstrate that this performance is maintained even in extreme low texture conditions. In summary, DRONE-SLAM delivers pose estimation accuracy comparable to GPS-RTK without its inherent limitations, such as dependency on signal availability and sensibility to obstacles.

References

[1] D. Navarro Tellez, E. Malis, R. Antoine, and P. Martinet, "Hybrid visual slam for multi-session precise localization: Application to a coastal cliff in Normandy", in *The Twenty-First International Conference on Autonomic and Autonomous Systems*, Mar. 2025, pp. 35–40, ISBN: 978-1-68558-241-8.

[2] D. Chabot, "Trends in drone research and applications as the journal of unmanned vehicle systems turns five", *Journal of Unmanned Vehicle Systems*, vol. 6, no. 1, pp. vi–xv, 2018.

[3] J. Fan and M. A. Saadeghvaziri, "Applications of drones in infrastructures: Challenges and opportunities", *International Journal of Mechanical and Mechatronics Engineering*, vol. 13, no. 10, pp. 649–655, 2019.

[4] R. Antoine *et al.*, "Geoscientists in the sky: Unmanned aerial vehicles responding to geohazards", *Surveys in Geophysics*, vol. 41, no. 6, pp. 1285–1321, 2020.

[5] A. Gupta and X. Fernando, "Simultaneous localization and mapping (slam) and data fusion in unmanned aerial vehicles: Recent advances and challenges", *Drones*, vol. 6, no. 4, p. 85, 2022.

[6] S. Halder and K. Afsari, "Robots in inspection and monitoring of buildings and infrastructure: A systematic review", *Applied Sciences*, vol. 13, no. 4, p. 2304, Jan. 2023, Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2076-3417. DOI: 10.3390/app13042304.

[7] Y. Ham, K. K. Han, J. J. Lin, and M. Golparvar-Fard, "Visual monitoring of civil infrastructure systems via camera-equipped Unmanned Aerial Vehicles (UAVs): A review of related works", *Visualization in Engineering*, vol. 4, no. 1, p. 1, Jan. 2016, ISSN: 2213-7459. DOI: 10.1186/s40327-015-0029-z.

[8] J. Jia and Y. Li, "Deep Learning for Structural Health Monitoring: Data, Algorithms, Applications, Challenges, and Trends", *Sensors*, vol. 23, no. 21, p. 8824, Jan. 2023, Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 1424-8220. DOI: 10.3390/s23218824.

[9] G. Esposito, A. Salari, I. Catapano, D. Erricolo, and F. Soldovieri, "UAV-based GPR prototype for structural monitoring of bridges: Preliminary results and perspectives", *ce/papers*, vol. 6, no. 5, pp. 930–933, 2023, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cepa.2074, ISSN: 2509-7075. DOI: 10.1002/cepa.2074.

[10] A. Massaro *et al.*, "Thermal IR and GPR UAV and Vehicle Embedded Sensor Non-Invasive Systems for Road and Bridge Inspections", in *2021 IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4.0&IoT)*, Jun. 2021, pp. 248–253. DOI: 10.1109/MetroInd4.0IoT51437.2021.9488483.

[11] A. Couturier and M. A. Akhloufi, "A review on absolute visual localization for UAV", *Robotics and Autonomous Systems*, vol. 135, p. 103 666, Jan. 2021, ISSN: 09218890. DOI: 10.1016/j.robot.2020.103666.

[12] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras", *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, Oct. 2016. DOI: 10.1109/tro.2017.2705103.

[13] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM", *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, Dec. 2021, IEEE Transactions on Robotics, ISSN: 1941-0468. DOI: 10.1109/TRO.2021.3075644.

[14] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time", in *2011 International Conference on Computer Vision*, Nov. 2011, pp. 2320–2327. DOI: 10.1109/ICCV.2011.6126513.

[15] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-Scale Direct Monocular SLAM", in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2014, pp. 834–849, ISBN: 978-3-319-10605-2. DOI: 10.1007/978-3-319-10605-2_54.

[16] B. Zhang and D. Zhu, "A Stereo SLAM System With Dense Mapping", *IEEE Access*, vol. 9, pp. 151 888–151 896, 2021, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3126837.

[17] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering", *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings)*, vol. 42, no. 4, Jul. 2023.

[18] R. A. Hamzah, R. A. Rahim, and Z. M. Noh, "Sum of Absolute Differences algorithm in stereo correspondence problem for stereo matching in computer vision application", in *2010 3rd International Conference on Computer Science and Information Technology*, vol. 1, Jul. 2010, pp. 652–657. DOI: 10.1109/ICCSIT.2010.5565062.

[19] J. Li *et al.*, "Practical stereo matching via cascaded recurrent network with adaptive correlation", in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, ISSN: 2575-7075, Jun. 2022, pp. 16 242–16 251. DOI: 10.1109/CVPR52688.2022.01578.

[20] P. S. Heckbert, "Fundamentals of Texture Mapping and Image Warping", University of California at Berkeley, USA, Technical Report, 1989.

[21] A. I. Comport, E. Malis, and P. Rives, "Real-time Quadrifocal Visual Odometry", *The International Journal of Robotics Research*, vol. 29, no. 2, pp. 245–266, 2010. DOI: 10.1177/0278364909356601.

[22] T. Junique *et al.*, "Investigation of the geological and hydrogeological structure of chalk cliffs with visible, thermal infrared and electrical resistivity imaging", *Journal of Hydrology*, vol. 630, p. 130 642, 2024.

[23] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles", in *Field and Service Robotics*, 2017. eprint: arXiv:1705.05065.

Finite-Word-Length-Effects in Mixed-Radix, Non-Power-of-2, Practical BFP-FFT

Gil Naveh
Toga Research Center
Huawei Technologies Co., Ltd
Hod-Hasharon, Israel
gil.naveh@huawei.com

Abstract— **Fixed-Point implementation of FFT is very sensitive to finite-word-length-effects due to the large quantization noise that is being accumulated throughout the FFT stages. In FFT implementations on fixed register size processors like CPUs and DSPs, Block-Floating-Point is a well-known scheme for controlling the tradeoffs between the fixed-point register size and the resultant accuracy. The performance of the radix-2 ideal Block-Floating-Point FFT, in terms of the output SQNR, has been investigated in depth. The ideal BFP-FFT suffers from implementation complexity, and especially non-deterministic latency. This results from the inherent mechanism which requires recalculating an entire FFT stage if one of that stage's outputs overflows. Because of this, most of the implementations are of a more practical variant of the BFP-FFT that does guarantee fixed latency. This, however, comes on the expense of reduced accuracy (degraded SQNR). In this paper we derive the SQNR formulas for the practical BFP-FFT for radix-2 and radix-4 Cooley-Tukey Decimation-In-Time FFTs, as well as for mixed-radix and non-power-of-2 FFTs. The derived model is compared to computer simulations and found highly accurate (less than 0.2 dB difference for the fixed radix, and less than 0.5 dB for the non-power-of-2 mixed radix). We use the derived model to compare the SQNR performance of the practical algorithm to the ideal one and show a 6-14 dB penalty for guaranteeing fixed latency implementation. For the mixed radix, the model enables to determine the optimal order of radices in terms of maximal SQNR at the FFT output.**

Keywords- Block-Floating-Point; Fixed-Point; DIT; SQNR; Mixed-Radix; Non-power-of-2.

I. INTRODUCTION

The Fast Fourier Transform (FFT) serves as an important tool in many signal processing applications. It has been successfully used in many applications such as radar, spectral analysis, filtering, voice processing and modems. Some of those are heavily relying on fixed-point processing. Since the FFT algorithm is known to be highly sensitive to finite-word-length effects (which are manifested as quantization noise), many attempts to derive an analytical model of the quantization noise have been conducted throughout the years. A rigorous such analysis for Decimation-In-Time (DIT) Block-Floating-Point FFT (BFP-FFT) for radix-2 and radix-4 is given in [1].

The vast majority of the applications and use-cases where FFT is being used require a power-of-2 FFT (FFT who's size, N, is a power of 2, e.g., 256, 512, 1024, etc.). This is the case in filtering via the convolution theorem, in DSL multitone modems [2], in wireless modems for multimedia [3], in fiber optics modems [4] and more. In the last few decades, a demand for mixed-radix, non-power-of-2 FFT has been showing up. Examples for this demand are cellular OFDM based modems like LTE, [5], and 5G-NR, [6].

In the cases that non-power-of-2 FFTs are required, when possible, one will extend the size to the next power-of-2 size and implement a power-of-2 FFT. However, there are cases where such extension is not possible, just like in LTE [5] or 5G-NR [6]. In those OFDM modems, there exist an uplink channel which relies on modulation scheme known as single-carrier OFDM. This modulation scheme is composed of two different FFT sections. In the first section the antenna data passes IFFT of sizes that are the product of $2^{m_1}3^{m_2}$ where $m_1 \geq 7$ and m_2 belong to the set $\{0,1\}$. In the second section the sequence of QAM modulated symbols is FFT transformed by a non-power-of-2 FFTs of sizes that are of the product $2^{m_1}3^{m_2}5^{m_3}$ where m_1, m_2 and m_3 are integers complying to $m_1 \geq 2, m_2 \geq 1, m_3 \geq 0$.

Finite-word-length effects have substantial implications on the accuracy performance of FFT. This is a result of the native characteristic of the FFT in which quantization noise that is added at the output of each stage of the FFT is accumulated toward the FFT output. Since the maximal value at each stage's output grows as we proceed with the stages [7], in many hardware implementations the performance degradation due to the quantization noise is mitigated by adapting the register size at each stage to accommodate the signal growth [8], [9], [10]. On the other hand, in Software implementations (as in CPUs and Digital Signal Processors - DSPs), or hardware implementations where intermediate values are forced to be written to memory, gradually increasing the bit-width of the stored values is not possible. For those cases, BFP based schemes are commonly used.

Among the BFP schemes, the dynamic-scale BFP lead to the highest accuracy for a given register size.

The straight-forward dynamic-scale BFP is such that throughout the calculation of each FFT stage, the butterflies' outputs are tested for an overflow. If an overflow is detected, the entire stage is recalculated and scaled down to prevent the overflow before being stored to memory. The advantage of this BFP scheme is that the scale down is done only on a concrete need, which leads to the best accuracy performance among other BFP-FFT algorithms. For that reason, we relate to the straight-forward dynamic-scale BFP-FFT as "ideal BFP-FFT" herein. The drawbacks of this algorithm are its complexity and the fact that that it results in non-deterministic latency. Deterministic latency may have high importance when the FFT is used within a synchronized pipelined system, such as a modulator or demodulator in OFDM modems [5].

An alternative BFP algorithm is such that there is a pre-defined down-scale factor at every stage [11]. This alternative has lower complexity and deterministic latency, but it's accuracy performance in terms of Signal-to-Quantization-Noise-Ratio (SQNR) is degraded as compared to the dynamic-scale BFP-FFT algorithms [12]. Another dynamic scale BFP scheme has been proposed by Shively [13]. In this scheme the decision of whether to scale down a certain FFT stage is determined as a function of the values of the outputs of the previous stage. That is, the decision whether to scale down a certain FFT stage is taken before the calculation of that stage is started, leading to a deterministic latency. This, on the other hand, comes on the expense of some loss in the FFT accuracy. Nevertheless, thanks to the deterministic latency of this scheme, it turns to be among the most commonly used schemes in practical implementations, e.g., [14], [15]. We refer to the Shively's scheme herein as "practical BFP-FFT". The original Shively's scheme aimed at Cooley-Tuckey, radix-2 FFT, and we extend it here to any Cooley-Tuckey, mixed-radix, power-of-2 and non-power-of-2 FFT.

The accuracy of non-BFP fixed-point FFT has been intensively analyzed as well as that of the pre-defined down-scale at every stage, e.g., [16]. The ideal BFP-FFT was originally analyzed in [17], which provided a lower and upper bound for the output quantization noise variance. In [7] and [12] a more accurate statistical model was used to project the expected value of the ideal BFP-FFT output noise power for an uncorrelated input sequence. A rigorous accuracy analysis of the practical BFP-FFT for power-of-2, fixed-radix DIT FFT is found in [1]. In the current paper we extend this analysis to mixed-radix and non-power-of-2 FFTs, where for power-of-2 we restrict ourselves to radix-2 and radix-4 (denoted as $\mathcal{R}2$ and $\mathcal{R}4$ hereafter) only. We derive an analytical model for the signal and noise power at the FFT output for any mixed-radix FFT by which the resultant SQNR can be predicted. Using the derived model one can also estimate the performance loss paid for using practical BFP as compared to an ideal BFP-FFT. For mixed-radix FFT, we show how the optimal order of radices of the given FFT size can be determined.

The problem of Twiddle Factors (TFs) quantization is not treated in this paper since the quantization effects of those are considerably lower than the computation roundoff errors [12].

The structure of the paper is as follows: in Section II the models of the underline FFT, the quantization, and the noise that are being used throughout the paper are defined. In Section III the SQNR formulas for a generic BFP-FFT scheme are derived. Section IV presents the scaling policies, and in Section V the SQNR formula for each of the scaling policies is provided. Section VI discusses the radices allocation throughout the FFT stages and the relations to the output SQNR for mixed-radix FFT. Results are presented in Section VII and the conclusions are given in Section VIII.

II. FFT, PROCESSOR AND QUANTIZATION NOISE MODELS

We relate to fixed-point representation of fractional datatypes. We assume a processor having registers of b bits (including sign) and accumulators of at least $B = 2b + \lceil \log_2 R \rceil + 1$ bits, where R is the FFT radix and $\lceil a \rceil$ is the smallest integer that is larger than a. The numbers represented by the registers are in 2's complement representation and in the range $-1 \le x \le 1 - 2^{-(b-1)}$. The numbers represented by the accumulators are in the range $-2^{\lceil \log_2 R \rceil + 1} \le x < 2^{\lceil \log_2 R \rceil + 1}$. The width of the data stored to memory is always of b bits. We assume complex multipliers that multiply complex multiplicands of b bits per component (b bits for the real component and b bits for the imaginary component). The output of the multiplier is of $2b + 1$ bits (as being a complex multiplication) that can be either scaled down and rounded to b bits, or added to an accumulator.

A generic scheme of a radix-R butterfly of DIT-FFT is given in Fig. 1. The inputs, x_n, are loaded from the memory and first multiplied by the Twiddle Factors (TFs), w_N^{kn}. After multiplication by the TFs, they are multiplied by the butterfly's coefficients $\gamma_{r,t}\,; r, t \in \{0, 1, \dots, R-1\}$, and then summed up within the butterfly to get the butterfly outputs, y_n, before being stored back to the memory. The processing model that we will deal here with, is a model that is most common to DSPs and dedicated FFT processors. In this model the inputs x_n and the TFs, w_N^{kn}, are represented by b bits per component (b bits for the real component and b bits for the imaginary component) and are within the range of $\left[-1\,, 1 - 2^{-(b-1)}\right]$. When multiplied, the multiplication is spanned over $2b + 1$ bits. In $\mathcal{R}2$ and $\mathcal{R}4$ FFTs the butterfly's

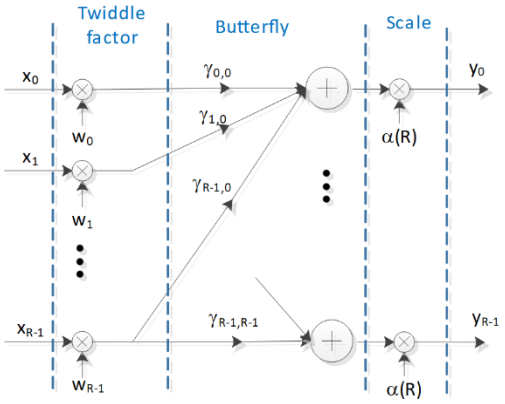


Fig. 1: Generic model of DIT FFT butterfly

internal coefficients, $\gamma_{r,t}$, belong to the sets $\{1,-1\}$ and $\{1,-1,j,-j\}$; $j=\sqrt{-1}$ respectively, and thus there are no actual multiplications within the butterfly. In those radices, the butterfly operation is, in fact, an addition or subtraction of the complex numbers or of their real-imaginary exchanged versions. This implies that the $2b+1$ bit-wide results of the multiplication by the TFs are not quantized before being summed up toward the butterfly output. The bit-width of the butterfly's output can grow and span over up to B bits and then potentially scaled down by a factor of α, where we restrict α to be of the form $\alpha=2^{-q}$ and q is a positive integer (the number of right shifts at the butterflies' outputs). The scaled down butterfly output is quantized to b bits per component, via rounding, before being stored to memory.

In radices other than $\mathcal{R}2$ and $\mathcal{R}4$ (i.e., non-power-of-2 radices), the butterflies' internal coefficients, $\gamma_{r,t}$, belong to the set $\left\{e^{-j\frac{2\pi rt}{R}}\right\}_{r,t=0}^{R-1}$, which implies that true complex multiplication takes place. Since the multiplier's multiplicands must be of b bits, the results of the TF multiplication are quantized to b bits before being multiplied by the butterfly internal coefficients.

The quantization model that we use here is the so-called Rounding-Half-Up (RHU) [18], which is also known as hardware-friendly-rounding and is being used in most digital signal processors and hardware implementations of digital signal processing functions. The mathematical function of RHU in rounding the value of s to b bits is

$$y=Q[s]\triangleq 2^{-b}\cdot\lfloor s\cdot 2^{b}+0.5\rfloor, \tag{1}$$

where $\lfloor a\rfloor$ is maximal integer lower than a and $s\in\left[-1,1-2^{-(b-1)}\right]$. The quantization error is $v=s-y$ and in the general case is modeled as an additive noise having uniform distribution [19]

$$v\sim U[-2^{-b},2^{-b}), \tag{2}$$

and is independent of s. As we deal here with finite-word-length, in fact, v has a discrete distribution. However, for large enough b, it is common to treat it as a continuous uniform distribution. We note also that by the definition of the RHU, v has an implicit bias since half way values of $s\cdot 2^{b}$ are always rounded up. Nevertheless, in most cases that s is of $2b$ bits, and b is large enough, this bias is negligible and hence the variance of the quantization noise is well approximated by the uniform RV variance

$$\sigma_v^2=\frac{2^{-2(b-1)}}{12}. \tag{3}$$

The model representing quantized values in $\mathcal{R}2$ and $\mathcal{R}4$ butterfly is given in Fig. 2. In this model there is a single quantization operation taking place at the butterfly output before being stored to memory. It is modeled as an additive noise source v, and we treat v as per (2). The model representing quantized values of butterflies with non-power-of-2 radices is given in Fig. 3. Here there is a noise source, v, modeling the quantization at the butterfly output, and a second noise source after the multiplication by the TF, u, that models the quantization noise caused by quantizing the result of the input multiplied by the TF.

In addition, throughout the FFT there are plenty of cases at which the butterfly's output value, before being scaled down and quantized, is a result of the summation of b-bits numbers multiplied by TF coefficients from the set

$$\mathcal{T}_1\triangleq\{1,-1,j,-j\}\ ;\ j=\sqrt{-1}\,, \tag{4}$$

i.e., all the coefficients toward a given butterfly output are among the set $\mathcal{T}_1$. We define those outputs as the set $\mathcal{O}$. In $\mathcal{R}2$ and $\mathcal{R}4$ butterflies, the outputs belonging to the set $\mathcal{O}$ are the outputs of butterflies that all the TFs preceding the butterfly belong the set $\mathcal{T}_1$. For the non-power-of-2 radices, the first output of any butterfly belongs to the set $\mathcal{O}$ (since $\gamma_{r,0}=e^{-j\frac{2\pi r\cdot 0}{R}}=1$ belong to the $\mathcal{T}_1$ set). In those cases, where all the coefficients toward a given butterfly output are among the set $\mathcal{T}_1$, the multiplication of a b-bits value $x\in\left[-1\,,1-2^{-(b-1)}\right]$ by the TF $w\in\mathcal{T}_1$ would result in a $2b$-bits number $a=w\cdot x$ that it's lower b bits are equal to zero. When such

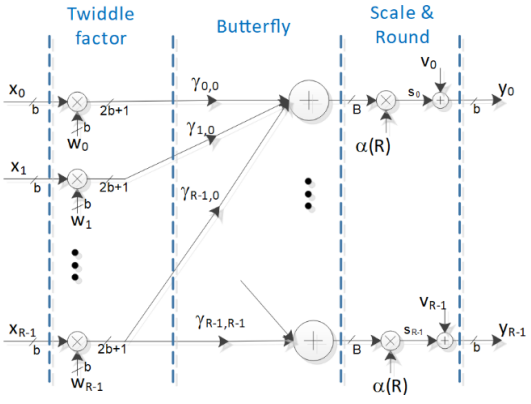


Fig. 2: $\mathcal{R}2$ and $\mathcal{R}4$ quantization noise butterfly model

a number is scaled down by very few bits, the quantization noise does not obey to the uniform distribution anymore [19]. In this case we get an RV having a discrete distribution and non-zero mean. For example, in the case that such a number is shifted one bit to the right, the quantization noise ε_1 is distributed as

$$\varepsilon_1 = \begin{cases} 0 & w.p.\ 0.5 \\ -\frac{1}{2}2^{-(b-1)} & w.p.\ 0.5\ , \end{cases} \tag{5}$$

where the subscript 1 in ε_1 refers to the case of quantization noise generated by right shift of the b-bits number by one bit. The expected value of this noise equals $-2^{-(b-1)}/4$ and hence, when dealing with SQNRs of those RVs, we will relate to the noise power rather than to its variance. To distinguish the power from the variance we use the symbol ρ^2 for power. The expected value of the power of ε_1 then is

$$\rho_{\varepsilon_1}^2 = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot\left(\frac{1}{2}2^{-(b-1)}\right)^2 = \frac{2^{-2(b-1)}}{8}. \tag{6}$$

As expected, this is larger than the variance of the zero mean uniformly distributed quantization noise of (3). In a similar way we can calculate the noise power of quantization noises that are generated due to the rounding after right shift of a b-bits number by q bits. In most FFT topologies and radices up to $\mathcal{R}5$, the right shifts are in the range of 0 to 3. Moreover, for right shifts of 4 and above, the quantization noise power is very close to the variance of the zero mean uniform quantization noise of (3). Therefore, for our analytical derivations we use

$$\rho_{\varepsilon_q}^2 = \begin{cases} 0 & ;\ q = 0 \\ \frac{1}{8}2^{-2(b-1)} & ;\ q = 1 \\ \frac{3}{32}2^{-2(b-1)} & ;\ q = 2 \\ \frac{11}{128}2^{-2(b-1)} & ;\ q = 3 \\ \frac{1}{12}2^{-2(b-1)} & ;\ q \geq 4\ . \end{cases} \tag{7}$$

III. SQNR OF A GENERIC BFP-FFT

By "generic BFP-FFT" we refer to a BFP-FFT that incorporates down-scaling by right shifts at the outputs of the FFT stages using an arbitrary scaling policy, where a scaling policy refers to the decision at which stages to scale down, and by what factor. For now, at which stages to scale down and by what factor will be parameters in the derivation. In the following paragraphs we will relate to specific BFP scaling policies and will analyze their SQNR performance. We assume zero mean i.i.d. input sequence, $x(n)$, and that the

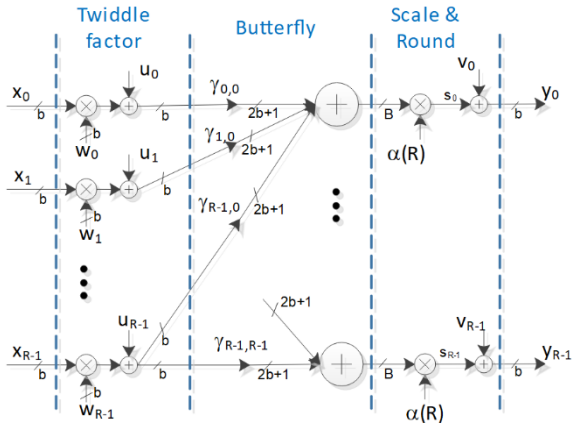


Fig. 3: Quantization noise model for Non-power-of-2 butterfly

quantization is regarded as an i.i.d. noise source. Moreover, multiple quantization noises at the input to a given butterfly that have been generated at earlier stages are mutually uncorrelated [12]. In order to derive the analytical expression of the SQNR, we will adopt the analysis strategy of Weinstein [12]. Let us relate to an input sequence of length N, $x(n)$, and a mixed-radix FFT with M stages. Denote the radix of the m^{th} stage as R_m such that $\prod_{m=1}^{M} R_m = N$. The scale value at the m^{th} stage is α_m, $m \in \{1, 2, \dots, M\}$, where we restrict α_m to be of the form $\alpha_m = 2^{-q_m}$ and q_m is the number of right shifts at the butterflies' outputs of the m^{th} stage. We denote $x_m(n)$ as the array values at the output of the m^{th} stage, where $x_M(k) \triangleq X(k)$ is the FFT output, and $x_0(n) \triangleq x(n)$ is the FFT input. For a zero mean, i.i.d. sequence $x(n)$, the variance of the signal at the FFT output is given by

$$\sigma_{x_M}^2 = N\sigma_{x_0}^2 \prod_{m=1}^{M} \alpha_m^2 = N\sigma_{x_0}^2 2^{-2\sum_{m=1}^{M} q_m}. \tag{8}$$

The noise at the output of a given butterfly is composed of two components: the noise that is generated by that particular butterfly, which we call butterfly self-noise, and the noise that is propagated through the butterfly (noise that was generated at earlier stages), which we call propagated-noise. At butterflies of $\mathcal{R}2$ and $\mathcal{R}4$, the self-noise is composed of a single noise source, v, at the butterfly output (refer to Fig. 2), while at the other, non-power-of-2, radices it is composed of the sum of R_m noise sources u_n, $n = 0 \cdots R_m - 1$, scaled down by α_m, plus a single v noise source at the butterfly output (refer to Fig. 3). Defining a uniform RV ξ distributed as $\xi \sim U[-2^{-b}, 2^{-b})$, and denoting the variance of the self-noise at each of the stage outputs as σ_B^2, we have

$$\sigma_B^2(m) = C_m \cdot \sigma_\xi^2, \tag{9}$$

where

$$C_m = \begin{cases} 1 & ; \quad R_m \in \{2,4\} \\ (R_m \alpha_m^2 + 1) & ; \quad R_m \notin \{2,4\}. \end{cases} \tag{10}$$

To simplify the description in the sequel, we define the set of radices $\mathcal{R}2$ and $\mathcal{R}4$ as the set $\mathcal{S}$.

The propagated-noise power passing through a butterfly is multiplied by a factor of $R_m \alpha_m^2$ as each butterfly output is composed of the sum of R_m i.i.d. input noise values and is multiplied by a scaling factor α_m. Looking at the propagated-noise at the output of an M stages FFT, it is observed that the self-noise from the first stage propagates through the following M-1 stages, which results in accumulation of $\prod_{m=2}^{M} R_m$ such i.i.d. noise sources, each attenuated by a factor of $\prod_{m=2}^{M} \alpha_m^2$. The propagation of the self-noise from the second stage results in accumulation $\prod_{m=3}^{M} R_m$ such i.i.d. noise sources, each attenuated by a factor of $\prod_{m=3}^{M} \alpha_m^2$, and so on. The total output noise variance, σ_E^2, for an M stages FFT, assuming all the quantization operations are modeled as uniform RVs, $U[-2^{-b}, 2^{-b})$, is therefore given by the following expression

$$\sigma_E^2 = \sigma_\xi^2 \left(C_M + \sum_{m=1}^{M-1} C_m \prod_{i=m+1}^{M} R_i \alpha_i^2 \right). \tag{11}$$

For the sake of simplicity of the formulation, we define a virtual $(M+1)^{th}$ stage at which $\alpha_{M+1} = \frac{1}{\sqrt{R_{M+1}}}$, and re-write (11) as

$$\sigma_E^2 = \sigma_\xi^2 \left(\sum_{m=1}^{M} C_m \prod_{i=m+1}^{M+1} R_i \alpha_i^2 \right). \tag{12}$$

In (11) and (12) it was assumed that the self-noise is a continuous RV and have the same PDF at all the outputs of all the butterflies. For b sufficiently large (e.g., $b = 16$) this assumption is commonly accepted. However, as explained in paragraph II, This is not the case for butterfly outputs of the set $\mathcal{O}$. The noise at the outputs of those butterflies is a discrete RVs and its Probability-Mass-Function (PMF) depends on the number of right shifts took place at the butterfly output. The power of those noise sources is larger than that of the zero-mean uniform RV, and hence they also have negative effect on the quantization noise power at the FFT output. In order to be able to evaluate the effect of those noise sources, we want to incorporate their statistical model in the derivation of σ_E^2 (or ρ_E^2). Before doing so, it worth mentioning two important notes. The first is that the distribution of the TFs of the set $\mathcal{T}_1$ among the FFT stages and among the butterflies within each stage is not uniform. Therefore, in each stage of radix $R \in \mathcal{S}$ there are some outputs that their self-noise has a non-uniform, non-zero-mean discrete PMF, and other outputs that their self-noise behaves as a continuous, zero-mean uniform RV. Similarly, in stages of radix $R \notin \mathcal{S}$, the first output of each butterfly is of the set $\mathcal{O}$, which has a non-uniform, non-zero-mean discrete PMF. The self-noise at the other outputs of those radices behaves as a continuous, zero-mean uniform RV. In addition, since each FFT output is connected (through the FFT flow graph) to a subset of the butterflies in each stage (except the first stage), the SQNR at the FFT output will not be identical at all output points. We will not relate to those effects here and will calculate the average SQNR at the FFT output sequence (average over all the output points). In fact, the noise power at the output of every stage of the FFT is not distributed evenly. But since we are interested in the average SQNR at the FFT output, we will also relate to the average noise power at the output of each stage of the FFT. The second note is the fact that the power of the sum of two non-zero-mean RVs does not equal to the sum of the powers like in two independent, zero-mean RVs as assumed in (12). However, since different noise sources are passing through different set of coefficients toward the same FFT output node, they can be assumed random and independent, justifying the use of the model of (12). There are very few FFT output nodes near the DC vicinity (near $k = 0$), that the set of coefficients along the path is correlated and the above assumption does not hold. Nevertheless, since the assumption does not hold only for a very small number of FFT output nodes, the effect on the overall averaged SQNR is negligible and the model of (12) can be used.

We denote by $\rho_{q_m}^2$ the noise power of a butterfly output noise source (noise source v) that belong to the set $\mathcal{O}$. The output noise power at those outputs is

$$\rho_O^2(m) = \begin{cases} \rho_{q_m}^2 & ; \quad R_m \in \mathcal{S} \\ (\sigma_\xi^2 R_m \alpha_m^2 + \rho_{q_m}^2) & ; \quad R_m \notin \mathcal{S}. \end{cases} \tag{13}$$

Denoting also by β_m the fraction of the outputs belonging to the set $\mathcal{O}$ at the m^{th} stage, we incorporate the effects of those outputs into the expression of the total output noise variance/power getting

$$\rho_E^2 = \sum_{m=1}^{M} [(1-\beta_m)\sigma_B^2(m) + \beta_m \rho_O^2(m)] \prod_{i=m+1}^{M+1} R_i \alpha_i^2. \tag{14}$$

Rearranging (14) and using (9), (10) and (13) we get

$$\rho_E^2 = \sum_{m=1}^{M} \left[C_m \sigma_\xi^2 + \beta_m \left(\rho_{q_m}^2 - \sigma_\xi^2 \right) \right] \prod_{i=m+1}^{M+1} R_i \alpha_i^2 . \tag{15}$$

The second term in (15), $\beta_m \left(\rho_{q_m}^2 - \sigma_\xi^2 \right)$, is a positive quantity that represents the increased output noise power caused by outputs of the set $\mathcal{O}$. The precise expression of β_m

as a function of the radix R can be extracted from the flow graphs of the FFTs. As stated before, for $\mathcal{R}2$ and $\mathcal{R}4$, outputs of the set $\mathcal{O}$ are caused by butterflies that all their preceding TFs are among the set $\mathcal{T}_1$, and for the non-power-of-2 radices, the first output of each butterfly belong to the set $\mathcal{O}$. The general rule is that at stages of non-power-of-2 radix, the fraction of the outputs of the set $\mathcal{O}$ is the reciprocal of the radix itself, i.e., R_m^{-1}, while for stages of radices $\mathcal{R}2$ or $\mathcal{R}4$, the fraction of outputs of the set $\mathcal{O}$ is one at the first stage ($m = 1$), and the product of the reciprocal of all preceding radices, $\prod_{i=1}^{m-1} R_i^{-1}$ for $m > 1$. Alternatively, this can be written as $R_m \prod_{i=1}^{m} R_i^{-1}$ for any m. An exception is the case that $R_m = 2$ and the radices of all preceding stages are among $\mathcal{S}$. In such a case the fraction of outputs of the set $\mathcal{O}$ is $2\prod_{i=1}^{m-1} R_i^{-1} = 4\prod_{i=1}^{m} R_i^{-1}$, $m > 1$. This is given by

$$\beta_m = \begin{cases} R_m^{-1} & ;\ R_m \notin \mathcal{S} \\ 4\prod_{i=1}^{m} R_i^{-1} & ;\ R_m = 2, \{R_i, i < m\} \in \mathcal{S} \\ R_m \prod_{i=1}^{m} R_i^{-1} & ;\ Otherwise. \end{cases} \tag{16}$$

Using (16) in (15), we can calculate the quantization noise at the FFT output, ρ_E^2. The output SQNR for a given scale pattern, $\boldsymbol{q} = [q_1, q_2, \dots, q_M]$, can be calculated, the by $\sigma_{x_M}^2/\rho_E^2$ from (8) and (15) respectively where assigning $\alpha_i = 2^{-q_i}$.

For a mixed-radix FFT, the output noise power of (15) is a function of the radices' distribution among the FFT stages. A precise expression for the output noise is a bit cumbersome. For fixed-radix FFTs, we can get a closed form for the output noise by introducing the expression of β_m into (15). For $\mathcal{R}2$ this results in

$$\begin{aligned} \rho_E^2 = \sigma_v^2 \sum_{m=1}^{M} R^{M-m+1} \prod_{i=m+1}^{M+1} \alpha_i^2 \\ +\left(\rho_{q_1}^2 - \sigma_\xi^2\right) R^M \prod_{i=2}^{M+1} \alpha_i^2 \\ + \sum_{m=2}^{M} \left(\rho_{q_m}^2 - \sigma_\xi^2\right) R^{M-2m+3} \prod_{i=m+1}^{M+1} \alpha_i^2 , \end{aligned} \tag{17}$$

for $\mathcal{R}4$ it results in

$$\begin{aligned} \rho_E^2 = \sigma_\xi^2 \sum_{m=1}^{M} R^{M-m+1} \prod_{i=m+1}^{M+1} \alpha_i^2 \\ + \sum_{m=1}^{M} \left(\rho_{q_m}^2 - \sigma_\xi^2\right) R^{M-2m+2} \prod_{i=m+1}^{M+1} \alpha_i^2, \end{aligned} \tag{18}$$

and for non-power-of-2, fixed-radix, in

$$\begin{aligned} \rho_E^2 = \sigma_\xi^2 \sum_{m=1}^{M} (R\alpha_m^2 + 1) \prod_{i=m+1}^{M+1} R\alpha_i^2 \\ + \sum_{m=1}^{M} \left(\rho_{q_m}^2 - \sigma_\xi^2\right) R^{-1} \prod_{i=m+1}^{M+1} R\alpha_i^2 . \end{aligned} \tag{19}$$

IV. Scaling Policies

Theoretically, one would like to pick a scaling policy that maximizes the Signal-to-Computation-Noise-Ratio of the finite-word-length FFT algorithm. Such maximization requires the allowance of overflows, which generates overload noise, and the optimization would be over the quantization plus overload noise. However, in most practical systems, such overflows are not allowed. As a result, the scaling policy is selected to maximize the SQNR under the constraint of zero-overflows. At the ideal BFP-FFT, the scaling policy is such that throughout the butterflies' computation, every butterfly's output is tested for an overflow before it is quantized down to b bits. If the real or the imaginary components of the butterfly output are smaller than -1.0 or larger than $1 - 2^{-(b-1)}$, the entire stage is re-calculated and the butterflies' outputs are scaled down by q bits before being rounded to b bits and stored to memory. The value q is selected to guarantee that the scaled result does not overflow anymore. For example, if one of the absolute values of the real or imaginary butterfly's outputs is within the range $\left[1\,, 2 - 2^{-(b-1)}\right]$, the entire stage will be re-calculated while the butterflies' outputs will be shifted by one bit to the right ($q = 1$). If one of the of the absolute values of the real or imaginary butterfly's outputs is within the range $\left[2\,, 4 - 2^{-(b-1)}\right]$, the entire stage will be re-calculated while the butterflies' outputs will be shifted by two bits to the right, and so on. As was mentioned in the introduction, this scheme suffers from non-deterministic latency and therefore is less favorable in practical implementations. The second, more common, policy is the one proposed by Shively [13], which guarantees deterministic latency and lower complexity at the expense of decreased SQNR. In this policy, the decision whether to down-scale the outputs of stage m and by what factor is taken based on the values of the outputs of stage $m - 1$, which are guaranteed to fit in the range $\left[-1\,, 1 - 2^{-(b-1)}\right]$. While writing the outputs of stage $m - 1$ to the memory, the processor finds the maximal absolute value among the real and imaginary components of the whole stage, and the down-scaling decision for the next stage is made according to this value. The down-scaling criterion is similar to the criterion being used by the scaling policy of the ideal BFP-FFT, i.e., to guarantee that no overflow will occur at the output of the next stage. Here, there is a need to consider the fact that the maximal absolute value at the butterflies' output of the m^{th}

stage would grow by a factor that is between 1 and $\sqrt{2}R_m$ relative to the outputs of stage $m-1$. In order to formalize this, let us define $x_m^c(n)$ for $n \in \{0, 1, \dots, N-1\}$ as

$$\begin{aligned} x_m^c(2n) &= real\big(x_m(n)\big) \\ x_m^c(2n+1) &= imag\big(x_m(n)\big), \end{aligned} \tag{20}$$

and

$$\tilde{x}_m = \max_n\{|x_m^c(n)|\}\,. \tag{21}$$

The scaling policy of the practical BFP-FFT can now be written as

$$q_m = \begin{cases} 0 & ;\ \tilde{x}_{m-1} < \dfrac{1}{\sqrt{2}R} \\ 1 & ;\ \dfrac{1}{\sqrt{2}R} \le \tilde{x}_{m-1} < \dfrac{2}{\sqrt{2}R} \\ 2 & ;\ \dfrac{2}{\sqrt{2}R} \le \tilde{x}_{m-1} < \dfrac{4}{\sqrt{2}R} \\ \vdots & \\ \lceil log_2(R) \rceil + 1 & ;\ \dfrac{1}{\sqrt{2}} \le \tilde{x}_{m-1} \end{cases}. \tag{22}$$

We denote the scaling policy of the ideal BFP-FFT as ϑ_i and of the practical BFP-FFT as ϑ_p.

V. SQNR CALCULATION

From the previous paragraph it is clear that the SQNR at the FFT output of a particular realization of the FFT depends on the scale pattern that has been used throughout this realization. Each scale pattern $\boldsymbol{q} = [q_1\,, q_2\,, \dots, q_M]$ is associated with a resultant SQNR. We adopt Weinstein's definition for "theoretical" SQNR as the weighted sum of the SQNR per scale pattern, i.e., the SQNR per scale pattern weighted by the probability of the particular scale pattern to occur [12]. The probability of a scale pattern depends on the radices allocation among the stages and the PDF of the input sequence. Of course, the radices allocation among the stages is a design parameter, therefore, for a given radices allocation, the probability of a scale pattern is solely dependent on the PDF of the input sequence and the scaling policy. In the sequel we will derive the scale patterns probabilities as well as the SQNR for the practical BFP-FFT algorithm and for the ideal BFP-FFT algorithm, for Gaussian input sequences. The Gaussian assumption simplifies the description, yet, the derivation can be adapted to any input sequence distribution.

A. SQNR of practical BFP-FFT

We start with the derivation of the probabilities of scale patterns. Given the practical BFP-FFT's scaling policy, the probability that there will be exactly $q > 0$ right shifts at stage m is equal to

$$\begin{aligned} &Pr\big(q_m = q; \vartheta_p\big) \\ &\quad = Pr\big(2^{q-1} \le \sqrt{2}R_m\tilde{x}_{m-1} \le 2^q\big) \\ &\quad = Pr\left(\frac{2^{q-1}}{\sqrt{2}R_m} \le \tilde{x}_{m-1} \le \frac{2^q}{\sqrt{2}R_m}\right) \\ &= Pr\left(-\frac{2^q}{\sqrt{2}R_m} \le \underset{n}{all}\{x_{m-1}^c(n)\} \le \frac{2^q}{\sqrt{2}R_m}\right) \\ &\quad -Pr\left(-\frac{2^{q-1}}{\sqrt{2}R_m} \le \underset{n}{all}\{x_{m-1}^c(n)\} \le \frac{2^{q-1}}{\sqrt{2}R_m}\right), \end{aligned} \tag{23}$$

whereas for $q = 0$

$$\begin{aligned} Pr\big(q_m = 0; \vartheta_p\big) &= Pr\big(\sqrt{2}R_m\tilde{x}_{m-1} \le 1\big) \\ &= Pr\left(\tilde{x}_{m-1} \le \frac{1}{\sqrt{2}R_m}\right). \end{aligned} \tag{24}$$

By the assumption that the input sequence, $x_{m-1}^c(n); n \in \{0, 1, \dots, 2N-1\}$ is an i.i.d. sequence, (23) and (24), can be written as

$$\begin{aligned} &Pr\big(q_m = q; \vartheta_p\big) \\ &\quad = \left[Pr\left(-\frac{2^q}{\sqrt{2}R_m} \le x_{m-1}^c(n) \le \frac{2^q}{\sqrt{2}R_m}\right)\right]^{2N} \\ &\quad -\left[Pr\left(-\frac{2^{q-1}}{\sqrt{2}R_m} \le x_{m-1}^c(n) \le \frac{2^{q-1}}{\sqrt{2}R_m}\right)\right]^{2N}, \end{aligned} \tag{25}$$

and

$$\begin{aligned} &Pr\big(q_m = 0; \vartheta_p\big) \\ &\quad = \left[Pr\left(-\frac{1}{\sqrt{2}R_m} \le x_{m-1}^c(n) \le \frac{1}{\sqrt{2}R_m}\right)\right]^{2N}. \end{aligned} \tag{26}$$

We now define the following auxiliary variables

$$T_m = 2^{-2Q_m}\,, \tag{27}$$

where

$$\begin{aligned} Q_m &= \sum_{i=1}^{m} q_i\ ;\ m \in \{1, 2, \dots, M\} \\ Q_0 &= 1\,, \end{aligned} \tag{28}$$

and

$$P_m = \prod_{i=1}^{m} R^i. \tag{29}$$

Using those, the variance of the sequence at the output of the m^{th} stage is

$$\sigma_{x_m}^2 = \sigma_{x_0}^2 P_m T_m, \tag{30}$$

and the variance of the real and imaginary individual components at the output of the m^{th} stage is $\sigma_{x_m}^2/2 = \sigma_{x_0}^2 P_m T_m/2$.

For an i.i.d. complex Gaussian input sequence, $x_0^c(n) \sim N\left(0, \sigma_{x_0}^2/2\right)$; $n \in \{0, 1, \dots, 2N-1\}$, it can be shown that all the intermediate sequences $x_m^c(n)$, $m \in \{1, 2, \dots, M\}$ are also Gaussian i.i.d. [12]. Therefore, the probability that the outputs of the m^{th} stage would be shifted by exactly $q > 0$ right shifts, given that there were accumulated Q_{m-1} right shifts at the stages preceding stage m is

$$\begin{aligned} &Pr\left(q_m = q \mid Q_{m-1}; \sigma_{x_0}^2, \vartheta_p\right) \\ &= \left[erf\left(\frac{\frac{2^q}{\sqrt{2}R_m}}{\sqrt{2}\frac{\sigma_{x_{m-1}}}{\sqrt{2}}}\right)\right]^{2N} - \left[erf\left(\frac{\frac{2^{q-1}}{\sqrt{2}R_m}}{\sqrt{2}\frac{\sigma_{x_{m-1}}}{\sqrt{2}}}\right)\right]^{2N} \\ &= \left[erf\left(\frac{2^q}{\sigma_{x_0}\sqrt{2P_m R_m T_{m-1}}}\right)\right]^{2N} \\ &\quad - \left[erf\left(\frac{2^{q-1}}{\sigma_{x_0}\sqrt{2P_m R_m T_{m-1}}}\right)\right]^{2N}, \end{aligned} \tag{31}$$

and the probability that there would be no right shifts ($q_m = 0$) is given by

$$\begin{aligned} &Pr\left(q_m = 0 \mid Q_{m-1}; \sigma_{x_0}^2, \vartheta_p\right) \\ &= \left[erf\left(\frac{\frac{1}{\sqrt{2}R_m}}{\sqrt{2}\frac{\sigma_{x_{m-1}}}{\sqrt{2}}}\right)\right]^{2N} \\ &= \left[erf\left(\frac{1}{\sigma_{x_0}\sqrt{2P_m R_m T_{m-1}}}\right)\right]^{2N}, \end{aligned} \tag{32}$$

where $erf(x)$ is defined by

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt\,. \tag{33}$$

We use those per-stage probabilities to calculate the probability of a specific scale pattern, $\boldsymbol{q} = [q_1, q_2, \dots, q_M]$,

$$\begin{aligned} &Pr\left(\boldsymbol{q}; \sigma_{x_0}^2, \vartheta_p\right) \\ &= Pr\left(q_1; \sigma_{x_0}^2, \vartheta_p\right) \prod_{m=2}^{M} Pr\left(q_m | Q_{m-1}; \sigma_{x_0}^2, \vartheta_p\right), \end{aligned} \tag{34}$$

and the output SQNR is calculated by the weighted sum of the SQNRs per scale pattern as

$$\begin{aligned} SQNR_{\vartheta_p} &= \sum_{\boldsymbol{q}} Pr\left(\boldsymbol{q}; \sigma_{x_0}^2, \vartheta_p\right) \cdot SQNR\left(\boldsymbol{q}, \sigma_{x_0}^2\right) \\ &= \sum_{\boldsymbol{q}} Pr\left(\boldsymbol{q}; \sigma_{x_0}^2, \vartheta_p\right) \cdot \frac{\sigma_{x_M}^2\left(\sigma_{x_0}^2\right)}{\rho_E^2\left(\boldsymbol{q}, \sigma_{x_0}^2\right)}. \end{aligned} \tag{35}$$

In (35) the expression $Pr\left(\boldsymbol{q}; \sigma_{x_0}^2, \vartheta_p\right)$ is calculated by (34), $\sigma_{x_M}^2\left(\sigma_{x_0}^2\right)$ is calculated by (8) and $\rho_E^2\left(\boldsymbol{q}, \sigma_{x_0}^2\right)$, with $\alpha_i = 2^{-q_i}$, is calculated by (17), (18) or (19) for $\mathcal{R}2$, $\mathcal{R}4$ and non-power-of-2 radices respectively. The number of different $\boldsymbol{q}$ patterns is quite large (e.g., for $\mathcal{R}2$, since q_m can take one of three options $\{0,1,2\}$ there are $3^{\log_2 N}$ optional different patterns). Nevertheless, the summation over all the $\boldsymbol{q}$ patterns in (35) can be calculated in reasonable time via a computer program.

Since we focus the analysis here on Gaussian inputs which are un-bounded in their values on one hand, while the FFT under analysis requires inputs in the range $\left[-1, 1 - 2^{-(b-1)}\right]$ on the other hand, we select the variance of the input signal such that the probability for values outside the allowed range at the input is sufficiently low. For the sake of the current analysis, we used $\sigma_{x_0} = 0.15$ which leads to a very low probability of having a sample outsize the allowed range. For example, for 4096 points FFT, the probability of having a vector of size 4096 with a sample outside the range $[-1,1]$ is approximately 10^{-7} (once per ten million FFT realizations, in average, there will be an input sample that has to be saturated to $\left[-1, 1 - 2^{-(b-1)}\right]$).

B. SQNR of the ideal BFP-FFT

At the scaling policy of the ideal BFP-FFT, ϑ_i, there are no pre-decisions for per-stage scaling. An FFT stage is calculated without scaling and throughout the calculations, if any of the stage's outputs overflows the allowed range, the whole stage is re-calculated while the outputs are down-scaled before being written to memory. Note that in the ideal policy there may be multiple re-calculation of the same stage if the strategy is to initiate the re-calculation upon the first overflowed value (strategy (a)). Different strategies that will eliminate the multi re-calculations of the same stage are: (b) upon the detection of the first overflow - set the scale value to the maximal scale value, and (c) always calculate the stage to its end and if overflows have been detected throughout the calculation, set the scale value according the largest magnitude among the detected overflowed values. Note that strategy (b) suffers degradations in the SQNR performance due to potential mismatch between the scale value and the actual maximal overflow value. Nevertheless, here, for the sake of SQNR comparison, we assume strategy (a) or (c), meaning that the scale is according to the largest magnitude output sample and no performance loss is involved. As opposed to the practical case where the scale decision for

stage m depends on $x_{m-1}(n)$, which are the outputs of stage $m-1$ after being scaled down, the scale decision of the ideal BFP-FFT depends on the output of stage m before being scaled down. Let us denote the output of stage m before being scaled down as $s_m(n)$, such that the scaled down values are

$$x_m(n) = \alpha_m s_m(n), \tag{36}$$

and define $s_m^c(n)$ and $\tilde{s}_m$ in analogous to (20) and (21) as

$$\begin{aligned} s_m^c(2n) &= real\big(s_m(n)\big) \\ s_m^c(2n+1) &= imag\big(s_m(n)\big), \end{aligned} \tag{37}$$

and

$$\tilde{s}_m = \max_n\{|s_m^c(n)|\}. \tag{38}$$

Now the SQNR analysis using the ideal BFP-FFT policy follows the steps of the analysis of the practical BFP-FFT scheme. The output signal variance and the output noise power follow (8) and (15), respectively. The probability that there will be exactly $q > 0$ right shifts at stage m is equal to

$$\begin{aligned} Pr(q_m = q;\vartheta_i) &= Pr(2^{q-1} \le \tilde{s}_m \le 2^q) \\ &= Pr\left(-2^q \le \underset{n}{all}\{s_m^c(n)\} \le 2^q\right) \\ &-Pr\left(-2^{q-1} \le \underset{n}{all}\{s_m^c(n)\} \le 2^{q-1}\right), \end{aligned} \tag{39}$$

and the probability that there will be no right shifts at stage m, i.e., $q = 0$, is

$$\begin{aligned} Pr(q_m = 0;\vartheta_i) &= Pr(\tilde{s}_m \le 1) \\ &= Pr\left(-1 \le \underset{n}{all}\{s_m^c(n)\} \le 1\right). \end{aligned} \tag{40}$$

Under the i.i.d. Gaussian input assumption, we get for $q > 0$

$$\begin{aligned} Pr\big(q_m = q \mid Q_{m-1};\sigma_{x_0}^2,\vartheta_i\big) &= \left[erf\left(\frac{2^q}{\sigma_{x_m}}\right)\right]^{2N} \\ &- \left[erf\left(\frac{2^{q-1}}{\sigma_{x_m}}\right)\right]^{2N} \\ &= \left[erf\left(\frac{2^q}{\sigma_{x_0}\sqrt{P_m T_{m-1}}}\right)\right]^{2N} \\ &- \left[erf\left(\frac{2^{q-1}}{\sigma_{x_0}\sqrt{P_m T_{m-1}}}\right)\right]^{2N}, \end{aligned} \tag{41}$$

and for $q_m = 0$

$$\begin{aligned} Pr\big(q_m = 0 \mid Q_{m-1};\sigma_{x_0}^2,\vartheta_i\big) &= \left[erf\left(\frac{1}{\sigma_{x_m}}\right)\right]^{2N} \\ &= \left[erf\left(\frac{1}{\sigma_{x_0}\sqrt{P_m T_{m-1}}}\right)\right]^{2N}. \end{aligned} \tag{42}$$

VI. RADICES ALLOCATION

For a mixed-radix FFT, the order of the radices (the allocation of radices to the various stages which forms a radices pattern) is a design parameter. Different orders will result in different scale pattern distributions and as a result - different output SQNR. In fact, the total amount of scaling (right shifts) of the ideal BFP-FFT for a given input realization depends solely on the values of the instantaneous input realization, and is independent of the order of radices. The number of right shifts in this case can be shown to be

$$Q_M = \left\lceil \log_2 \max_k(|Real\{X(k)\}|, |Imag\{X(k)\}|) \right\rceil, \tag{43}$$

where $X(k)$ is the FFT output for the specific input realization, assuming no scaling take place throughout the FFT. At the practical BFP-FFT the total number of down scaling is not completely independent on the order of the radices. It depends on the radix allocated to the last stage, stage M, and is in the range $\{Q_M, Q_M+1, \cdots, Q_M + \lceil \log_2(\sqrt{2}R_M)\rceil\}$.

The output noise, on the other hand, does depend on the scaling patterns, while those depend on the order of the radices. The variance of the resultant SQNR between various radices-patterns is not large and is shown to be in the range of 0.2 dB to 2.25 dB for the LTE DFT sizes. An easy way to determine the best order of radices is to calculated the SQNR (according to (35)) for all the radices permutations and pick the one with the highest SQNR. In Fig. 4 the best and worst SQNR among all the radices permutations for each of the LTE DFT sizes is shown. An interesting observation from Fig. 4 is that for the non-power-of-2, mixed-radix FFT of the LTE sizes, the SQNR is not necessarily a monotonic function of the FFT size. As can be seen there is an average monotonicity, but not local monotonicity. The reason is the fact that in close sizes, despite the fact that the size is close, the set of radices involved is different. Since the quantization noise generated by a butterfly of non-power-of-2 radix is larger than that of a butterfly of power-of-2 radix (refer to (9) and (10)), an FFT that involves more non-power-of-2 radices, is likely to result in larger output quantization noise. For example, the sizes of 324, 360 and 384 are three consecutive sizes in Fig. 4, and show monotonic increasing SQNR. When examining the radices involved, we find that size 324 include four stages of non-power-of-2 radices (since $324 = 4 \cdot 3^4$), size 360 include three stages of non-power-of-2 radices ($360 = 4 \cdot 2 \cdot 5 \cdot 3^2$), and the size 384 include only one radix which is non-power-of-2 ($384 = 2 \cdot 4^3 \cdot 3$).

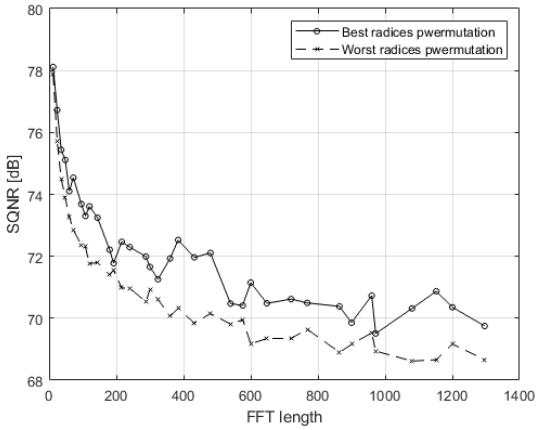


Fig. 4: SQNR of best and worst radices permutations for non-power-of-2 FFTs of LTE sizes

VII. RESULTS

The derived models of the SQNR of the practical and the ideal BFP-FFT have been validated against simulation. The model and the simulation results for 16-bit datatype ($b = 16$) and Gaussian i.i.d. input with standard deviation of $\sigma_{x_0} = 0.15$ are shown in Fig. 5 and Fig. 6 for radix-2 and radix-4 respectively. The simulation result vs. the BFP model for non-power-of-2, mixed radix, practical BFP-FFT of the LTE sizes is shown in Fig. 7. For the simulation results we have averaged the SQNR of 1000 FFT runs per FFT length. As can be seen, there is a very good match between the simulation results and the derived model in all cases. The gap between the refined statistical model (that incorporate the refinement for butterfly outputs of the set $\mathcal{O}$) and the simulation result for the practical BFP-FFT is in the order of 0.2 dB for the fixed-radix, power-of-2 FFTs and in the order of 0.5 dB for the mixed-radix, non-power-of-2 FFTs. The simulation results for the ideal BFP-FFT are not shown in the figures since the model has almost perfect match to the simulation result with gaps that are in the order of 0.05 dB.

In Fig. 5 and Fig. 6 we can also see the effect of the refined statistical model for the butterfly outputs of the set $\mathcal{O}$. In Fig. 5 is it seen that the model neglecting the effects of the butterfly outputs of the set $\mathcal{O}$, for radix-2 BFP-FFT, is optimistic by about 0.5 dB for the practical BFP-FFT and in Fig. 6 it is optimistic by about 1 dB for radix-4.

One of the main goals of the paper is to provide an analytical tool that enables the prediction of the SQNR penalty one needs to pay for getting fixed latency BFP-FFT. This penalty is clearly seen for radix-2 and radix-4 in Fig. 5 and Fig. 6 respectively. We see that such a penalty is in the order of 6 dB when the number of stages is above five, and grows up to 13.5 dB for lower number of stages as seen at the case of 64 points radix-4 FFT. The reason that for low number of stages the degradation of the practical BFP-FFT is larger, is the fact that the difference between the number of truly required down-scales (used by the ideal BFP-FFT) and the number of down-scales used by the practical BFP-FFT (Shively's scheme) reduces as the number of stages grows and that in the practical BFP-FFT the scaling take place at earlier stages.

Another interesting observation that the model reveals relates to the comparison of the SQNR between radix-2 and radix-4 BFP-FFT implementations for a power-of-2 fixed-radix FFT. It is well known that from complexity perspective, the radix-4 has advantages over radix-2 (at least in the number of multiplications). From the results in Fig. 5 and Fig. 6, we can also see that radix-4 have better SQNR in the ideal BFP-FFT implementation. We get 4 dB advantage for 64-points FFT down to about 2 dB advantage for 4096-points FFT. However, for the practical BFP-FFT we see an opposite behavior. The radix-2 practical BFP-FFT results in 2.8 dB better SQNR for 64-point FFT, down to 1.2 dB better SQNR for 4096-points FFT. The reason for this phenomenon is that the number of the quantization noise sources depends on the number of stages, such that in the radix-4 FFT there are half the number of noise sources as compared to radix-2, while the total down-scaling depends on the type of the BFT-FFT. For ideal BFP-FFT the total down scaling of radix-2 and radix4 is the same (as given in (43)). Hence, since radix-2 has more quantization sources, it also has lower SQNR performance as compared to radix-4. For the practical BFP-FFT, number of down-scaling of the radix-2 and radix-4 FFTs may not be the same. Since the maximal absolute value is a monotonic, non-decreasing, function of the stage index (it always non-decreasing between consecutive stages) [7], the number of down-scales of the practical BFP-FFT would be greater or equal to that of the radix-2. As a result, the signal power at the output of the radix-4 practical BFP-FFT is lower or equal that that of the radix-2 and hence, despite the fact that there are more noise sources in radix-2 the total SQNR is better

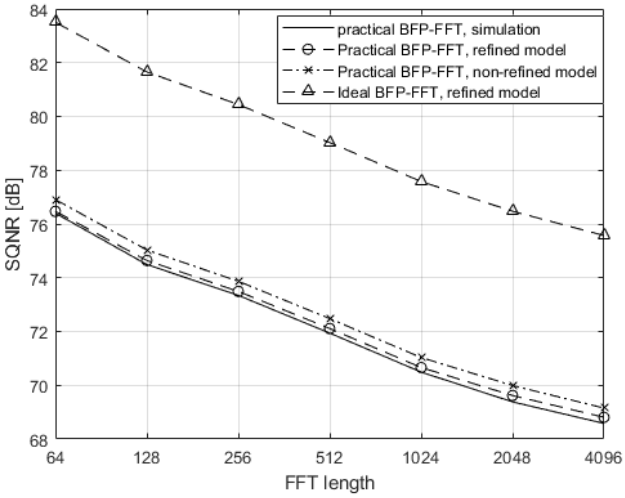


Fig. 5: Radix-2 Practical BFP-FFT

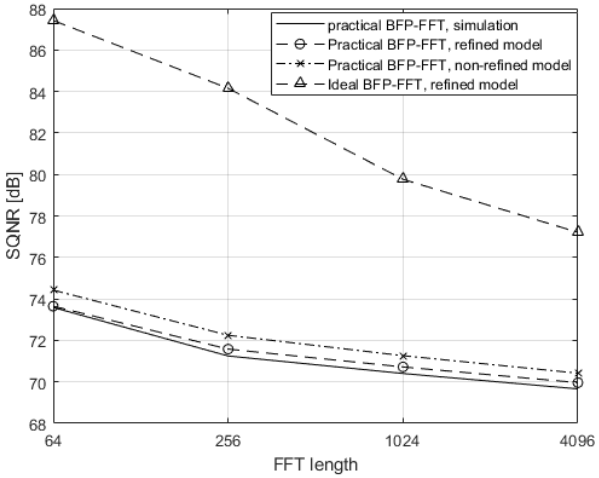


Fig. 6: Radix-4 Practical BFP-FFT

VIII. CONCLUSIONS

In this paper we extended the analytical model of the finite-word-length-effects of Cooley Tukey DIT BFP-FFT of [1] to cover fixed-radix, as well as mixed-radix, non-power-of-2 FFTs. We incorporate butterfly outputs belonging to the $\mathcal{O}$ set as a refined model, and derived the analytical expressions for the ideal and practical BFP-FFTs. The models have been validated against simulation and found highly accurate for both, the ideal and the practical BFP-FFTs. The model enables to accurately predict the SQNR for the practical BFP-FFT and the performance degradation compared to the ideal BFP-FFT scheme. The model also can be used to determine the best radix order of mixed-radix FFTs as described in paragraph VI.

The derivation covers DIT-FFT and refer to a straight-forward implementation model of non-power-of-2 butterfly. The framework used can be easily adapted to other topologies and other implementation models of the non-power-of-2 butterflies.

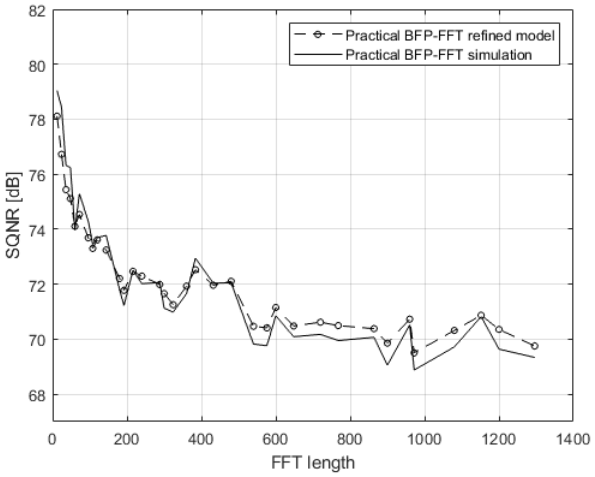


Fig. 7: Mixed-Radix LTE sizes Practical BFP-FFT

REFERENCES

[1] G. Naveh, "Finite-Word-Length-Effects in Practical Block-Floating-Point FFT," in *SIGNAL 2025*, Lisbon, 2025.

[2] J. M. Cioffi, V. Oksman, J. J. Werner, T. Pollet, P. Spruyt, J. S. Chow and K. S. Jacobsen, "Very-high-speed digital subscriber lines," *IEEE Communications Magazine*, vol. 37, no. 4, pp. 72-79, 1999.

[3] B. F. Frederiksen and R. Prasad, "An overview of OFDM and Related Techniques Towards Development of Future Wireless Multimedia Communications," in *IEEE Proc. Radio and Wireless Conference*, Boston, 2002.

[4] N. Cvijetic, "OFDM for Next-Generation Optical Access Networks," *IEEE Journal of Lightwave Technology*, vol. 30, no. 4, pp. 384-398, 2012.

[5] *LTE-A; Evolved Universal Terrestrial Radio Access (E-UTRA), Physical Channels and Modulation*, 3GPP TS 36.211, 2011.

[6] *NR; Physical Channels and Modulation*, 3GPP TS 38.211, 2025.

[7] A. V. Oppenheim and C. J. Weinstein, "Effects of Finite Register Length in Digital Filtering and the Fast Fourier Transform," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 957-976, 1972.

[8] W.-H. Chang and N. Q. Truong, "On the Fixed-Point Accuracy Analysis of FFT Algorithms," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 4973-4682, 2008.

[9] P. Gupta, "Accurate Performance Analysis of a Fixed Point FFT," in *Twenty Second National Conference on Communication (NCC)*, Guwahati, 2016.

[10] A. Monther and K. Zsolk, "Analysis of Quantization Noise in FFT Algorithms for Real-Valued Input Signals," in *International Conference on Radioelektronika*, Kosice, 2022.

[11] S. Qadeer, M. Z. Ali Khan and S. A. Sattar, "On Fixed Point error analysis of FFT algorithm," *ACEEE Int. Journal on Information Technology*, vol. 01, no. 03, 2011.

[12] C. J. Weinstein, "Quantization Effects in Digital Filters," M.I.T. Lincoln Lab. Tech. Rep. 468, ASTIA doc. DDC AD-706862, 1969.

[13] R. R. Shively, "A Digital Processor to Generate Spectra in Real Time," *IEEE Transactions on Computers*, Vols. C-17, no. 5, pp. 485-491, 1968.

[14] H. G. Kim, K. T. Yoon, J. S. Youn and J. R. Choi, "8K-point Pipelined FFT/IFFT with Compact Memory for DVB-T using Block Floating-point Scaling Technique," in *International Symposium on Wireless Pervasive Computing (ISWPC)*, Melbourde, 2009.

[15] S. J. Huang and S. G. Chen, "A High-Parallelism Memory-Based FFT Processor with high SQNR and novel addressing scheme," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, Montreal, 2016.

[16] Tran-Thong and B. Liu, "Fixed-Point Fast Fourier Transform Error Analysis," *IEEE Transactions on Acoustic, Specch, and Signal Processing*, vol. 24, no. 6, pp. 563-573, 1976.

[17] P. D. Welch, "A Fixed-Point Fast Fourier Transform Error Analysis," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 2, pp. 151-157, 1969.

[18] L. Xia, M. Athonissen, M. Hochstenbach and B. Koren, "Improved Stochastic Rounding," *arXiv , 2020, Available: https://arxiv.org/abs/2006.00489.*

[19] B. Widrow, I. Kollar and M.-C. Liu, "Statistical theory of Quantization," *IEEE Transactions on Instrumentation and Measurement*, vol. 45, no. 2, pp. 353-361, 1996.

Differential Power Amplifiers in 130 nm Partially Depleted and 28 nm Full Depleted Silicon-On-Insulator Technologies for 5G Applications

[1]Marcos L. Carneiro, [2]Tristan Lecocq, [3]Eric Kerhervé, [4]Magali de Matos, [5]Thierry Taris, [6]Jean-Marie Pham
[1]Postgraduate Program in Industrial and Systems Engineering, Pontifícia Universidade Católica de Goiás, Goiânia, Brazil.
[1,2,3,4,5,6]University of Bordeaux, CNRS UMR 5218, Bordeaux INP, Talence, France.
e-mail: [1]mcarneiro@pucgoias.edu.br, [2]tristanlecocq@free.fr, [3]eric.kerherve@bordeaux-inp.fr, [4]magali.de-matos@u-bordeaux.fr, [5]thierry.taris@ims-bordeaux.fr, [6]jean-marie.pham@u-bordeaux.fr

Abstract— **This paper starts with a review of the 5G Narrow Band-Internet of Things (NB-IoT) and it provides an analysis of the 130 nm Partially Depleted (PD) Silicon-On-Insulator (SOI) and 28 nm Full Depleted (FD) SOI technologies. It proposes the design of two Power Amplifiers (PAs) for 5G NB-IoT applications and presents performance graphs for S-parameters, Power-Added Efficiency (PAE), gain, output power (*Pout*), frequency sweep, and different biasing setups for the back-gate voltage. Both PAs consist of a gain stage (driver) and a power stage, using pseudo-differential and cascode topologies. The 28 nm PA includes an additional stacked transistor in the power stage to accommodate a higher drain bias voltage. They were fabricated and measured, demonstrating the gain adjustment capability of FDSOI technology via back-gate voltage, which allowed for approximately 3.6 dB of gain adjustment. Both PAs met the required performance parameters in post-layout simulations, achieving maximum Power-Added Efficiency (*PAEmax*) of 49% and 38.5%, gain of 36 dB and 34 dB and saturated Power (*Psat*) of 32 dBm and 28.8 dBm, respectively for 130 nm and 28 nm, placing them at the state-of-the art.**

Keywords- Power Amplifier; CMOS; 130 nm PDSOI; 28 nm FDSOI; 5G applications; Nb-IoT.

I. INTRODUCTION

This paper is an extended version of [1]. This version adds a review section about the Narrow Band-Internet of Things (NB-IoT), the integration technologies of 130 nm Partially Depleted (PD) Silicon-On-Insulator (SOI) and 28 nm Full Depleted (FD) SOI, and Power Amplifiers (PA). In the design methodology section, the resistivity of the thick metal layers as a function of the track width for 130 nm PDSOI and 28 nm FDSOI technologies was added, along with details about the designed interstage wideband transformer used in both technologies and the cascode active balun used in the 28 nm circuit. On results it was added the 28nm FDSOI PA Pout vs Pin measured performances for 3 levels of back-gate voltage and the Pout and PAE performances, related to the frequency between 1 GHz and 3 GHz, for circuits on both technologies.

The transition from 4G Long Term Evolution (LTE) to 5G has revolutionized the Internet of Things (IoT) with the advent of massive IoT, enabling the connection of numerous devices simultaneously. Narrow Band-Internet of Things (NB-IoT), a key 5G standard within Low-Power Wide-Area Networks (LPWAN), addresses the need for massive IoT by supporting battery-powered devices with extended lifespans and optimized installation costs. Operating on licensed 3GPP bands, NB-IoT offers higher data rates compared to unlicensed LPWAN technologies like Long Range (LoRa) and Sigfox. It achieves extensive coverage through transmission repetitions and increased signaling power, while its Single-Carrier Frequency Division Multiple Access (SC-FDMA) modulation reduces Peak-to-Average Power Ratio (PAPR), improving Power Amplifier (PA) efficiency and ensuring suitability for massive IoT applications [2].

Silicon-on-insulator (SOI) technology is pivotal for overcoming RF integration challenges in IoT circuits. Leveraging the high integration capabilities of Complementary Metal-Oxide-Semiconductor (CMOS), SOI reduces parasitic capacitances with a BOX layer, enhancing performance by over 20% [3]. While SOI improves reliability, energy efficiency, and reduces variability compared to bulk CMOS [4], NB-IoT's Single-Carrier Frequency-Division Multiple Access (SC-FDMA) modulation imposes strict PA design requirements, demanding linear operation and efficiency at low power. Advanced SOI technologies like Partially Depleted SOI (PDSOI) and Full Depleted SOI (FDSOI) provide tailored solutions, excelling in isolation and low-power scenarios, respectively [5].

This paper analyzes the 130 nm PDSOI and 28 nm FDSOI technologies, and it proposes the design of two PAs for the 5G NB-IoT applications (see Fig. 1). The gain and linearity adjustment capability via the back-gate voltage of FDSOI technology is demonstrated. Both circuits consist of PAs with a gain stage (driver) and a power stage, using pseudo-differential and cascode topologies.

Section II presents a review of the NB-IoT, the integration technologies of 130 nm PDSOI and 28 nm FDSOI, and PAs classes. Section III compares 130 nm PDSOI and 28 nm FDSOI technologies, highlighting their components and PA design methodology. Section IV presents post-layout simulation and measurement results, including performance analysis, gain tuning via back-gate voltage for the 28 nm PA, and a state-of-the-art comparison. Section V concludes with findings and future research directions.

II. TECHNOLOGY AND CIRCUIT REVIEW

With the evolution from 4G to 5G and the growing demand for connected devices, developing new standards that allow an increase in the number of simultaneously connected

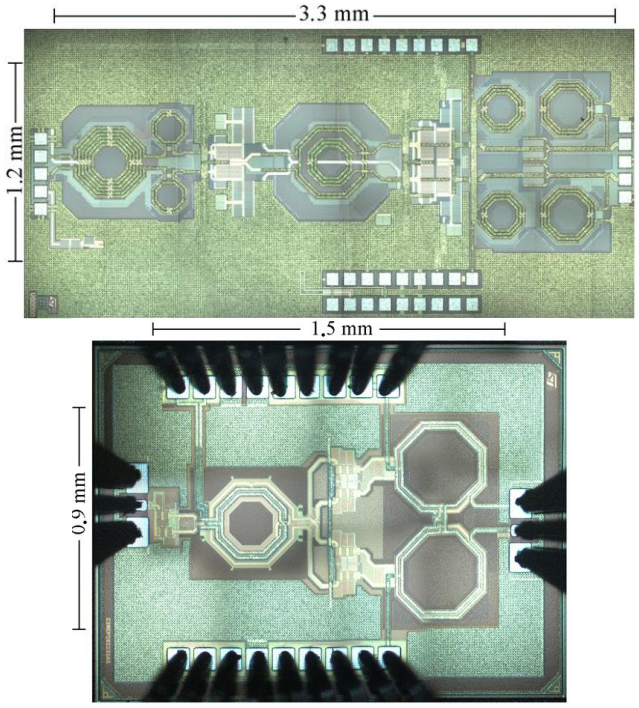


Figure 1. 130nm PDSOI PA (top) and 28nm FDSOI PA (bottom).

devices is essential. This section is dedicated to presenting the NB-IoT standard, the benefits of using silicon technologies for IoT circuits, and topics on the characterization and design of power amplifiers (PAs), as well as the key parameters that affect design.

A. NB-IoT Standard

The IoT refers to a network of electronic devices that gather, process, and communicate environmental information, such as temperature and pressure. Devices, or IoT nodes, transmit data to a base station, which centralizes information for analysis. Data can then be stored and processed in the Cloud or locally, depending on whether the base station's computing capacity is sufficient [6]. IoT networks can be organized as local area networks (LANs) for limited geographical areas, connecting to the internet via wireless protocols such as Wi-Fi or physical cables for improved efficiency and speed. For broader coverage, wide area networks (WANs) with low-power wireless standards (LPWANs) are employed. LPWANs enable battery-powered IoT devices to operate for extended periods, thereby optimizing installation costs by reducing the number of required base stations. Standards such as LoRa and Sigfox utilize free ISM bands, enabling cost-free operation with limited data rates. In contrast, NB-IoT and Cat-M operate in licensed 3GPP bands, which support higher data rates but incur higher costs due to licensing [2].

The 5G network, as the fifth generation of cellular communication standards, builds upon advancements in speed, bandwidth, and functionality that have evolved since the inception of mobile networks in the 1980s with 1G. Initially limited to analog voice transmission, each generation has expanded capabilities to meet growing demands, with 4G in 2016 facilitating support for IoT. The deployment of 5G not only continues these improvements but also targets modernization in industrial sectors and the development of smart cities [7]. This evolution, along with the rapid growth in the industrial IoT market [8], calls for revisiting existing IoT standards. Massive IoT, a concept introduced with 5G, is designed to support a high density of connected devices through enhanced LPWAN technologies and the capabilities of 5G. Its defining features include high connection density, irregular non-critical data transfers, small data packet sizes, and stringent energy efficiency requirements, aiming for over 10 years of battery life [9].

To meet the stringent requirements of massive IoT, 5G introduces two dedicated standards, Cat-M and NB-IoT. This paper focuses on NB-IoT, due to its extensive research applications. NB-IoT, an LPWAN standard, reuses 4G LTE frequency bands and protocols to streamline deployment. This standard emphasizes extended network coverage and ultra-low device complexity [10]. Coverage improvements are achieved through transmission repetitions and increased signaling power, with three repetition modes allowing varying degrees of coverage enhancement. However, increased repetition and power can impact energy consumption, requiring a trade-off between low power use and extensive coverage. To reduce device complexity, NB-IoT optimizes the physical layer, lowering the computing demands for signal processing. It also employs SC-FDMA modulation for the uplink, which reduces PAPR and improves the power amplifier's efficiency [11]. NB-IoT encompasses multiple emission classes (e.g., classes 3, 5, and 6) to adjust power usage according to quality of service (QoS) requirements, thereby optimizing energy consumption [12]. Additionally, energy-saving techniques such as eDRX and PSM are employed. eDRX allows for the periodic shutdown of the receiver, while PSM enables deep sleep mode by turning off the radio module for negotiated periods, further extending battery life [13].

SC-FDMA modulation is crucial in minimizing energy consumption for NB-IoT devices, as it enables single-carrier characteristics through a digital Fourier transform (DFT) while maintaining subcarrier allocation, thereby supporting multiple connections [14], [15]. Unlike OFDMA, where each QPSK symbol occupies one subcarrier for the entire symbol duration, SC-FDMA encodes each QPSK symbol across all N subcarriers for 1/N of the symbol duration, achieving lower PAPR. This flexibility enables simultaneous multi-user access on the same LTE channel and dynamically allocates resource blocks (RBs) according to user demand, with two modes of allocation—localized and interleaved. For example, in a 10 MHz LTE channel, there are 50 RBs, each with 12

subcarriers, resulting in 600 addressable subcarriers for allocation [14], [15]. SC-FDMA is well-suited for massive IoT applications by balancing energy efficiency and user capacity within LTE constraints [14].

B. SOI Technology for RF function integration in Silicon

CMOS SOI technology involves manufacturing a wafer with an inserted insulating layer, resulting in a silicon-insulator-silicon substrate stack. The thickness of the upper silicon layer can range from 5nm to several micrometers. Various insulators are used for these wafers (e.g., sapphire, silicon oxide), with silicon oxide being the most common for low-cost applications. The most widely used SOI wafer fabrication processes are Separation by IMplanted OXygen (SIMOX) and the Smart-Cut method, which together account for 90% of SOI wafer production [16].

The SIMOX process begins with implanting a large amount of oxygen into a standard wafer. A high-temperature annealing step then transforms the oxygen ions into a silicon oxide layer. The oxide layer's thickness and depth are controlled by annealing temperature, dose, and energy used during oxygen ion implantation [16].

While advanced silicon-on-insulator (SOI) technologies such as SIMOX improve silicon's electrical performance, silicon-based materials still face limitations in high-power applications. As a result, power amplifiers for commercial use—particularly in mobile telephony—often rely on III-V technologies, such as GaAs or GaN, which offer superior power performance compared to silicon technologies (see Fig. 2). Silicon has lower breakdown voltages, making it less competitive with these materials. Additionally, CMOS has a lower maximum operating frequency than its III-V counterparts, making it easier to design power amplifiers with III-V technologies. However, CMOS offers high integration capability and low cost, attracting industrial interest, especially in massive IoT applications.

CMOS SOI technology inherits many advantages from bulk silicon technology, particularly its integration capability and low manufacturing cost. SOI technology shares many aspects of the CMOS fabrication process, thereby keeping production costs low. Adding the steps to convert a CMOS wafer into a CMOS SOI wafer only increases costs by about

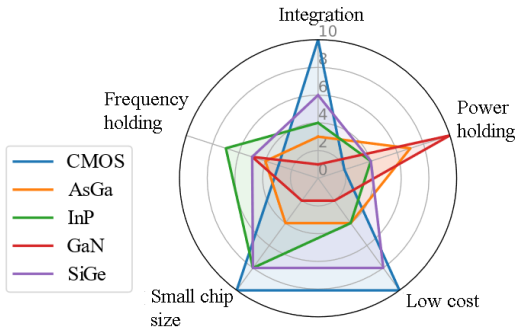


Figure 2. Comparing integrated technologies.

10% to 15%, making SOI competitive in terms of cost [4]. Additionally, the cost gap narrows as technology nodes shrink, providing a significant advantage for SOI.

SOI technology enhances performance by over 20% due to the reduction of parasitic capacitances in transistors [3]. This improvement primarily results from isolating the transistor's active area from the substrate with a BOX layer, thereby preventing the direct connection of PN junction capacitances to the transistor and reducing parasitic capacitances. SOI also offers improved reliability, reduced process complexity and variability, and lower energy consumption compared to CMOS technology [4].

Both CMOS and CMOS SOI technologies continue to evolve, particularly with decreasing minimum feature sizes, resulting in enhanced transistor frequency performance, improved integration, and lower circuit production costs. However, transistor miniaturization also increases their susceptibility to degradation due to the reduction in maximum supported voltages for each junction.

To enhance integration, technologies are adding more metal layers, which reduces the individual layer thickness and thus the maximum current density. Lower maximum voltages and current densities further limit the power handling of CMOS technologies, so several degradation mechanisms must be considered in PA design [17] as gate oxide breakdown, hot carrier injection, punch-through effect, floating body effect, and electromigration.

C. Power Amplifiers

The power amplifier is the circuit responsible for amplifying the input signal to a specified output power level, as defined by the NB-IoT standard, before transmitting it to the antenna. It is therefore subject to constraints related to power, linearity, efficiency, and thermal management.

Power amplifiers are categorized into two main operating classes: linear (or sinusoidal) and switching. Linear classes, including Class A, AB, B, and C, are defined by the transistor's conduction time, set by its biasing, where the transistor operates as a controlled current source [18]. This mode yields a linear relationship between the input and output power, making it suitable for signals that require linearity.

Switching classes, such as Class D, E, F, and G, are based on the harmonic treatment of the output network or the input signal processing. Here, the transistor functions as a switch, and these classes are generally suitable only for constant-envelope signals. However, since NB-IoT employs SC-FDMA modulation, which does not have a constant envelope, only linear operating classes are suitable for this application.

The choice of an operating class for a PA depends on efficiency, output power, gain, and linearity constraints. As the conduction angle decreases, power gain also decreases due to the increased excursion of the input signal required to reach the maximum current. Operating classes A, AB, and B

provide a balance between efficiency, gain, and linearity.

As modulation complexity and/or operating frequency increase, the amplifier must be more linear and have maximum gain, guiding the choice toward Class A.

For NB-IoT, with a frequency below 6 GHz, gain and linearity must still be optimized, making a deep Class AB—close to Class B—the preferred choice.

III. Design Methodology

This design methodology section presents a study on technologies and the designed PAs. Details about the metal layers of the 130 nm FDSOI and 28 nm PDSOI technologies are presented, followed by a comparison of the inductors, capacitors, and transistors of both technologies. The parameters are presented in the order in which the designer should analyze them during the project. Based on this analysis, the following subsection provides details of the schematics of the two designed PAs, highlighting their similarities and differences.

A. Evaluation of Passives and Transistors of SOI Technology

Fig. 3 presents the metal layers of the 130 nm PDSOI and 28 nm FDSOI. The first observation concerns the difference in the number of available metal layers and their thickness. Indeed, the smaller the technology node, the higher the integration density, which also requires an increase in interconnection density. Several solutions have been implemented to increase this density [19]. The rise in metal layers and the reduction of the minimum etching widths are the most common and easiest to apply. However, reducing the minimum etching width impacts the maximum thickness of metal layers that can be achieved due to manufacturing processes. This consequently explains the reduction in the thickness of the metal layers in the 28 nm FDSOI.

To determine the quality of the interconnections, the graphs presented in Figure 4 show the resistivity of the thick metal layers as a function of the track width. The metallization of the 130 nm PDSOI exhibits significantly improved performance due to the thicker layers of both aluminum (ALU, LB) and copper (M4U, Ix). The curves are plotted between the W_{min} and W_{max} of each level. Thus, in addition to having higher resistivity, the 28 nm FDSOI also has lower maximum widths in Ix layers. This must be considered in electromigration calculations to ensure that the conductor is sufficiently wide to carry the desired current. The densification of the metal layers also leads to a decrease in the maximum voltages between two metal levels due to the phenomenon of Time-Dependent Dielectric Breakdown (TDDB) [20]. Consequently, the maximum power supported by the passives is reduced.

Figures 5 and 6 illustrate the performance of inductors and capacitors from each technology, respectively. The

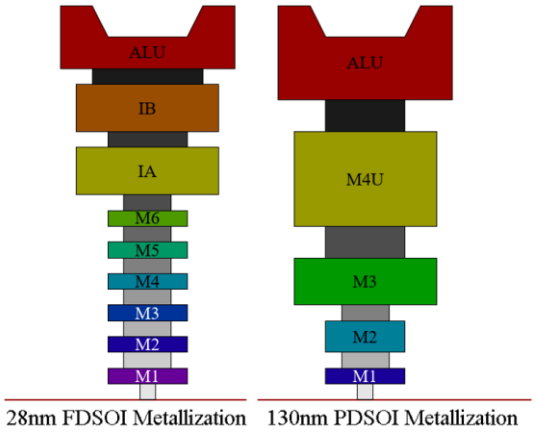


Figure 3. Metal layers of 28 nm FDSOI and 130 nm PDSOI technologies

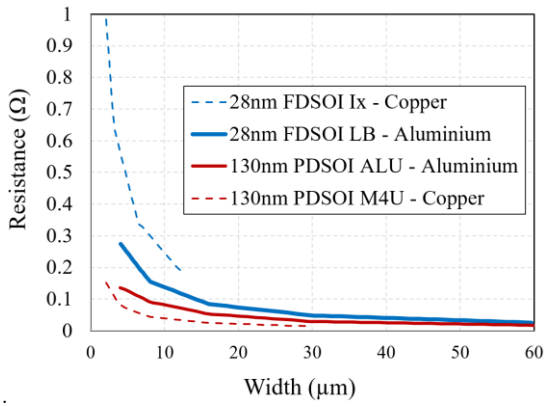


Figure 4. Resistivity of the thick metal layers as a function of the track width for 130 nm PDSOI and 28 nm FDSOI technologies.

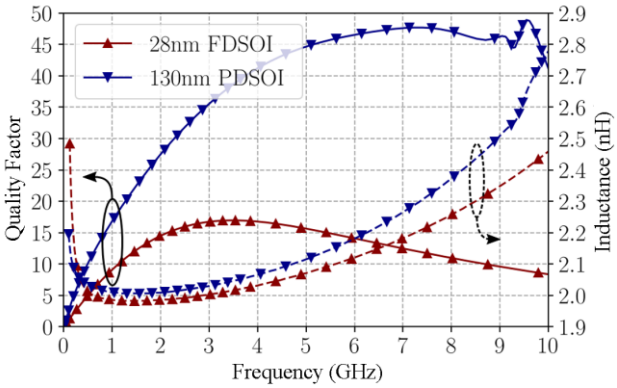


Figure 5. Comparison of inductances from 28 nm FDSOI and 130 nm PDSOI technologies.

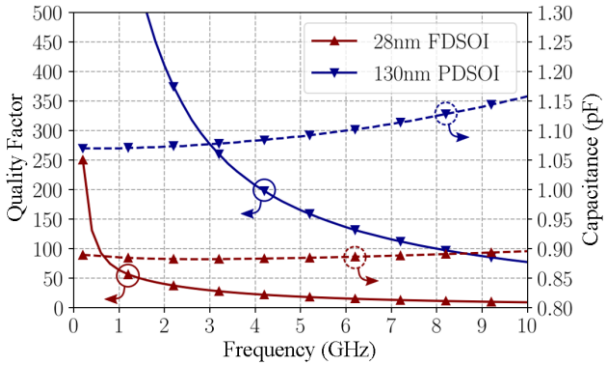


Figure 6. Comparison of capacitances in 28 nm FDSOI and 130 nm PDSOI technologies.

comparison was made with inductors using an octagonal topology [20]. In 28 nm FDSOI, the inductors are designed on the three thick levels, ALU-IB-IA (see Fig. 3), to reduce resistivity and increase the quality factor at low frequencies. In 130 nm PDSOI, two thick metal levels, ALU-M4U (see Fig. 3), are utilized. For the same topology, the inductor achieves a quality factor *Q* of 28 at 2 GHz in 130 nm PDSOI, compared to 15 in 28 nm FDSOI. However, high-value inductors exhibit better high-frequency behavior in 28 nm FDSOI due to a higher self-resonant frequency, indicating lower parasitic capacitances. For capacitors, the quality factor at 2 GHz in the 130 nm technology is approximately 300 for a capacitance of 1.1 pF (see Fig. 6). In contrast, the 28 nm technology yields a quality factor of 40 at 2 GHz for a capacitance of 0.88 pF. Indeed, the 28 nm technology has much thinner and more resistive metal layers than the 130 nm technology. On the other hand, the capacitors in 130 nm occupy larger silicon areas.

Figures 7 and 8 show the output transfer characteristics of NMOS transistors for RF applications in PA design. The transistors from 28 nm FDSOI have a higher current density, reaching 1.2 mA at the maximum *Vgs* voltage, compared to 0.58 mA for the thick oxide transistor in 130 nm PDSOI. Additionally, the 28 nm transistors have lower threshold voltages, around 250 mV, compared to approximately 350 mV for the 130 nm transistors, enabling operation at lower voltages.

The 130 nm PDSOI technology offers improved transistor quality in the saturation region. Indeed, the slopes *∂Id/∂Vds* in the saturation region are lower for the 130 nm PDSOI transistors than for the 28 nm FDSOI transistors. This also indicates that the *gds* in 130 nm is lower than in 28 nm. The consequence is achieving more linear transistors for large-signal applications.

In summary, the electrical characteristics of the 28 nm FDSOI transistors—higher current density, lower threshold voltage, and reduced operating voltage—make them more suitable for energy-efficient, compact, and long-range IoT RF transmitters, particularly in systems requiring LPWAN or massive-IoT operation. These advantages enable smaller chips, longer battery life, and more reliable communication in constrained environments.

B. Power Amplifier Design Methodology

The two PA architectures were designed (see Figs. 9 and 10) based on preliminary transistor sizing and analysis of the presented passive components. Both architectures were designed to achieve comparable performance and NB-IoT restrictions in post-layout simulations. This allows for evaluating their fabricated circuit measurements to compare the two technologies and discuss their advantages and limitations in relation to the target application.

Each circuit includes a driver stage with a single-ended input and pseudo-differential cascode topology at the output. Additionally, both circuits feature a pseudo-differential cascode power stage. The 130 nm PDSOI PA, depicted in Fig. 9, incorporates a pseudo-differential cascode power stage alongside a pseudo-differential cascode driver setup. This configuration ensures a straightforward design and excellent

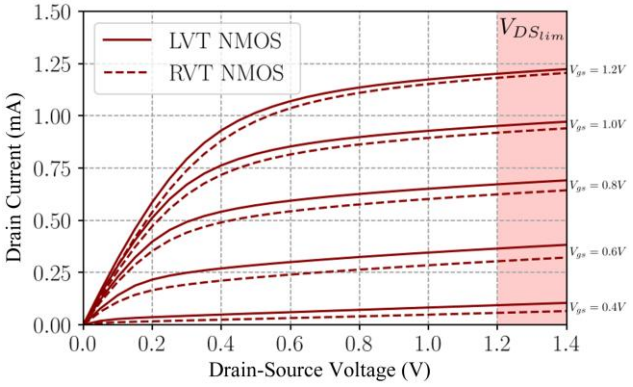


Figure 7. Output Transfer Characteristics Id(Vds) in 28 nm FDSOI.

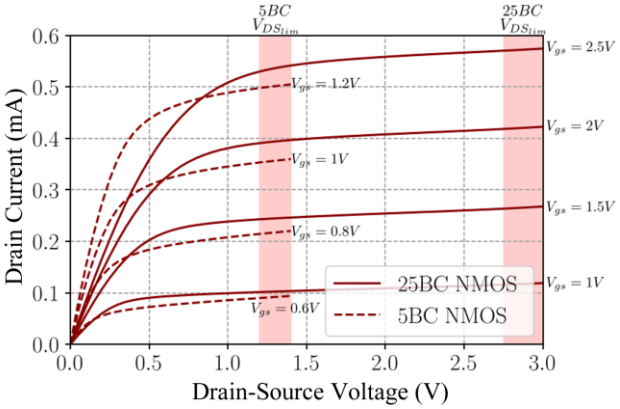


Figure 8. Output Transfer Characteristics Id(Vds) in 130 nm PDSOI.

performance tailored for NB-IoT applications. The design achieves higher output power by employing pseudo-differential architecture while minimizing constraints on the ground return path by suppressing even harmonics. Furthermore, the cascoded transistor arrangement enhances the amplifier's gain, allowing it to meet the 35 dB target specification. To ensure stability, given the high gain, neutralization capacitors (C_{neutro}) are incorporated. The matching networks are designed to enable broadband operation facilitated by a broadband matching transformer. In the 130 nm technology, the power stage transistors were dimensioned with W_{total} = 1200 µm, and the driver stage transistors were dimensioned with W_{total} = 300 µm. The circuit was biased with Vdd = 5 V.

The interstage wideband transformer in the 130nm PA (see Fig. 11a), located between the driver and the power cell, was designed to present a coupling coefficient (k) of 0.35 at the central frequency of 1.85 GHz. The simulated optimal conjugated output impedance of the driver is R^*_{optDRV} = 50 − 50j Ω and the input impedance of the power cell is R_{inPC} = 5 − 27.2j Ω. The wideband transformer presents inductances of 3.4 nH and 2.8 nH in the primary and secondary, respectively. The quality factors are 2.8 and 5.1 for the primary and secondary, respectively.

The transformer exhibits an average insertion loss of 2.1 dB, with a minimum of 1.9 dB at 1.54 GHz. The significant losses are mainly due to the reduced coupling coefficient of the transformer and the quality factors of the

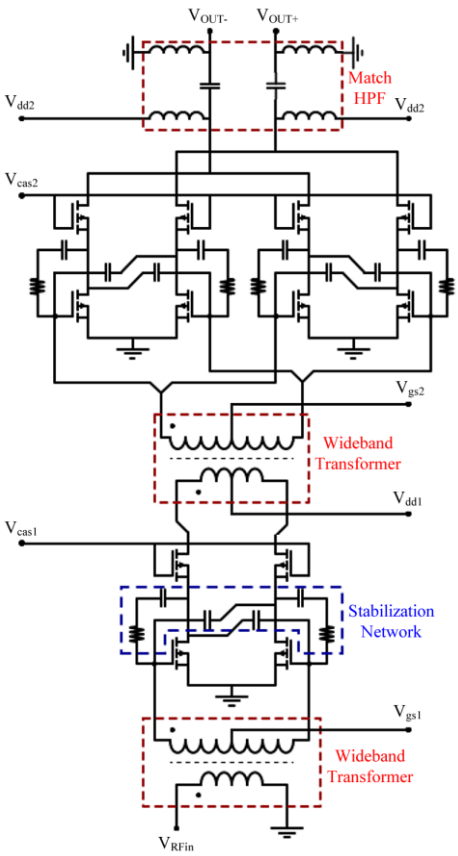


Figure 9. PA in 130 nm PDSOI technology schematic.

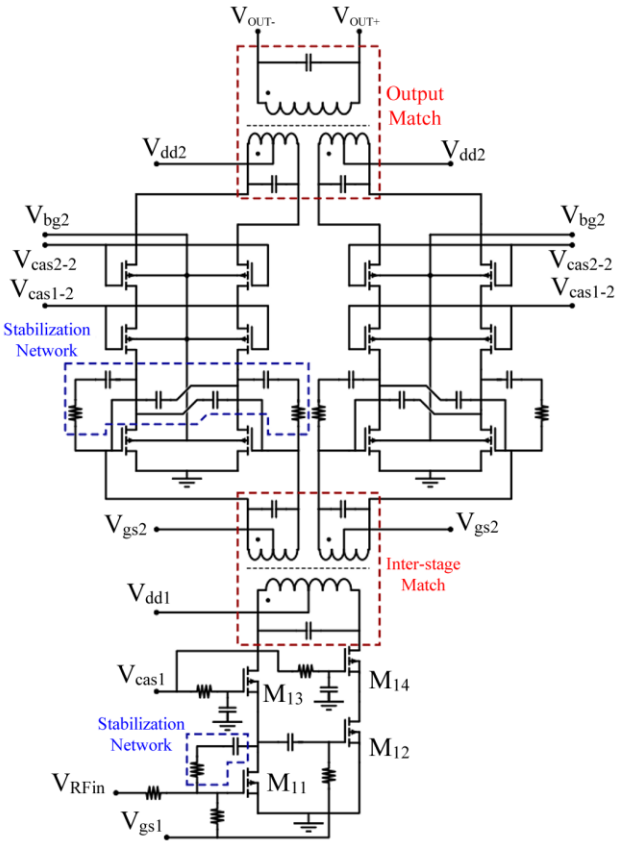


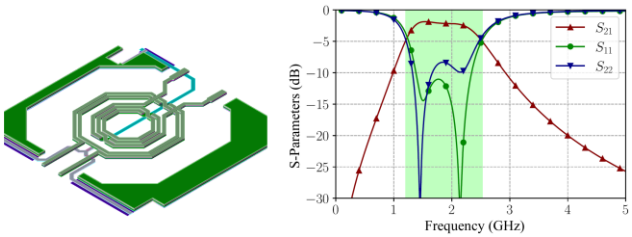
Figure 10. Schematic PA in 28 nm FDSOI technology schematic.

inductors in 130 nm PDSOI technology. The 3 dB bandwidth of the circuit is 1.3 GHz, with a lower cutoff frequency at 1.2 GHz, allowing coverage of the entire targeted NB-IoT band. Figure 11b presents the transformer's small signal performance.

Figure 10 shows the complete schematic of the PA in 28 nm FDSOI technology. This design utilizes two power cells with differential triple-stack cascode topology in its power stage to enable a supply voltage (Vdd) closer to the 130 nm technology, facilitating a more accurate comparison. The power cells are combined to compensate for the technology's power limitations, enabling a total output power of 28 dBm.

The output matching network uses a distributed active transformer (DAT) to optimize the load impedance at the output through series recombination. The inter-stage matching is designed around a 2-to-4 transformer (Figure 12), which performs impedance matching while distributing power across each power cell. Finally, the driver employs a cascode active balun topology [21], eliminating the need for a passive input balun. In the 28 nm technology, the power stage used transistors with W_{total}=900 µm, and for the driver stage, W_{total}=225 µm, and the circuit was biased with Vdd=3V.

The cascode active balun in the input of the 28 nm PA functions as an amplifier that transforms a single-ended input into a differential output, while also supplying sufficient gain for the subsequent power stage [5]. This approach helps minimize chip area by eliminating the need for a passive balun



(a) Transformer layout (b) S-Parameters performance

Figure 11. (a) Transformer layout in the 130 nm PA and (b) small signal performance.

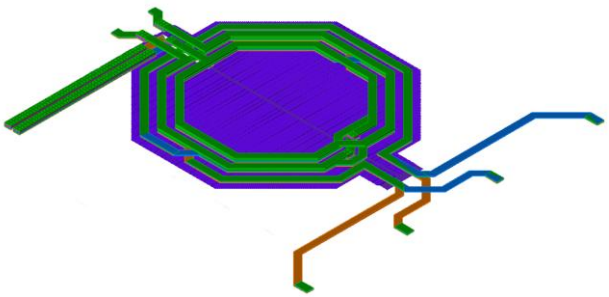


Figure 12. Inter-stage matching 2-to-4 transformer layout in the 28nm PA.

typically placed before the driver stage. The active balun design (shown in Fig. 9) employs a common-source transistor (M11) to provide the initial 180° phase shift. The resulting signal is then directed through a second common-source transistor (M12), introducing an additional 180° shift, effectively producing a signal in phase with the original input. Both differential signals are further amplified using common-gate transistors (M13 and M14), which serve as cascode stages to enhance gain, output isolation, and power capability. A series capacitor connected to the gate of M12 serves a dual purpose: it decouples the transistor's DC bias and allows fine-tuning of the signal level reaching M12, ensuring a balanced output power across both branches.

This circuit explores the potential for improving output power by utilizing a stacked architecture and back-gate biasing, aiming to meet the power requirements of NB-IoT applications. The back-gate voltage allows for fine-tuning of the gain and linearity performance, as demonstrated in the results section.

In the case of inductors and transformers, the coplanar topology was preferred in 130 nm PDSOI due to the limited number of thick metal layers. In 28 nm FDSOI, stacked transformers are favored because of their better coupling factor. Both circuits were designed to achieve post-layout simulations (PLS) at a central frequency of 1.85 GHz, with a bandwidth exceeding 400 MHz, a gain of 35 dB, a maximum Power-Added Efficiency (PAE_{max}) above 30%, and a power back-off PAE (PAE_{PBO}) above 20%.

IV. Results and Discussions

A. Post-Layout Simulation and Measurement Performance

The S-parameter performance of the PA in 130 nm PDSOI and the PA in 28 nm FDSOI technologies post-layout simulation (PLS) and measurement from 1 GHz to 3 GHz are presented in Figs. 13 and 14, respectively. The 130 nm PA presents an almost constant *S21* performance (between 35 dB and 39 dB) from 1.55 GHz to 2.4 GHz, an *S22* near -3 dB, and an *S11* less than to -5 dB in this frequency range. The 28 nm PA presents flatter behavior, with a maximum *S21* performance of 33 dB between 1.5 GHz and 1.8 GHz, *S22* less than -5 dB, and *S11* less than -15 dB.

The gain and PAE performances for the 130 nm PA from post-layout simulation and measurements at a frequency of 1.85 GHz are presented in Fig. 15. The measured gain performance exhibits a class AB characteristic shape, with a gain of 34.5 dB at low power and a maximum of 36 dB. The maximum PAE reaches 48.5% at a P_{sat} of 31 dB in PLS and 38% in measurements at a P_{sat} of 28 dBm.

The gain and PAE performances for the 28 nm PA from post-layout simulation at a frequency of 1.85 GHz are presented in Fig. 16. The gain performance achieves 33.26 dB in low power and a maximum of 34.72 dB; the maximum PAE reaches 38.5% at a P_{sat} of 28.5 dB. The transistors were optimized until the edge of stability was reached, predicting that losses in further components would ensure stability. However, the implemented circuit presented stability issues at high output power.

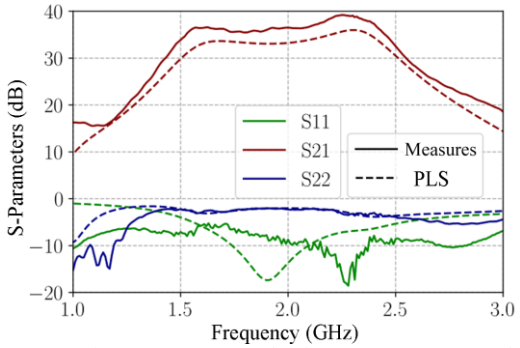


Figure 13. PA in 130nm PDSOI technology S-parameters post-layout simulation (PLS) and measurement performance.

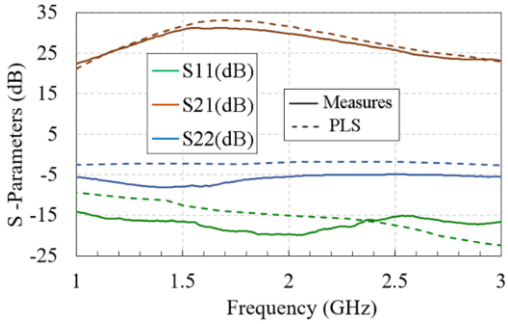


Figure 14. PA in 28nm FDSOI technology S-parameters post-layout simulation (PLS) and measurement performance.

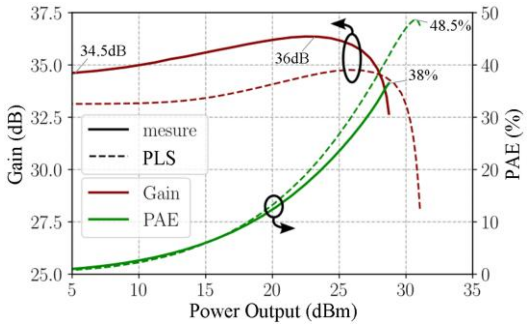


Figure 15. 130nm PA gain, PAE post-layout simulation, and measured performances in 1.85 GHz.

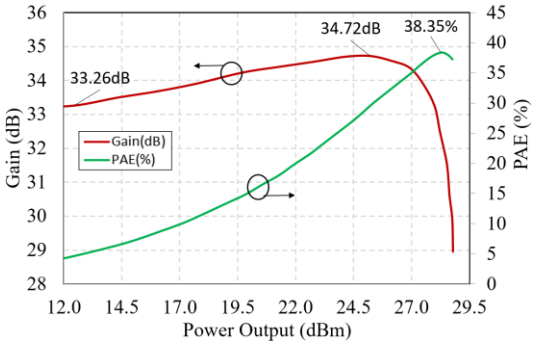


Figure 16. 28nm PA gain and PAE post-layout simulation performance in 1.85 GHz.

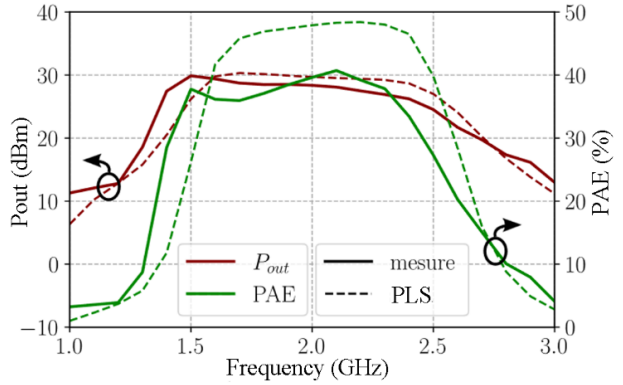


Figure 17. 130 nm PA Pout and PAE performances between 1 GHz and 3 GHz.

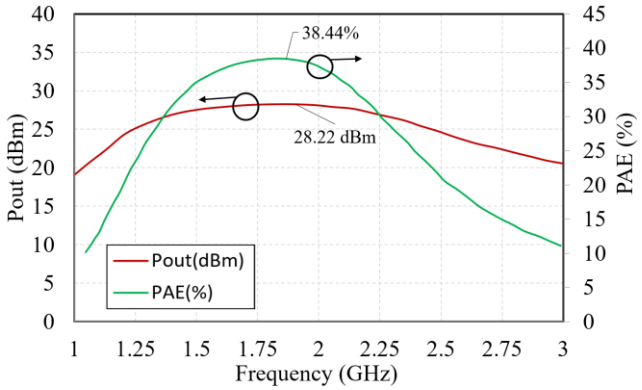


Figure 18. 28 nm PA Pout and PAE performances (PLS) between 1 GHz and 3 GHz.

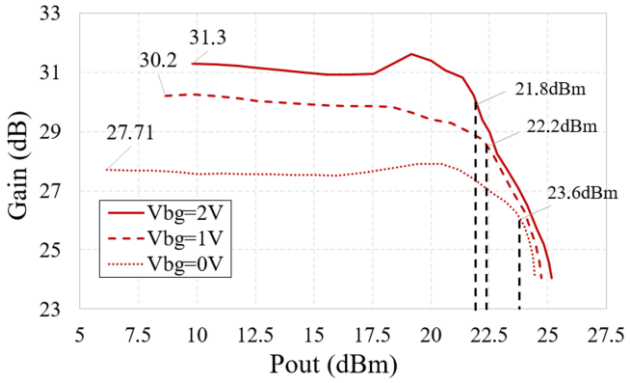


Figure 19. 28nm FDSOI PA gain versus P_{out} measured performances for 3 levels of back-gate voltage.

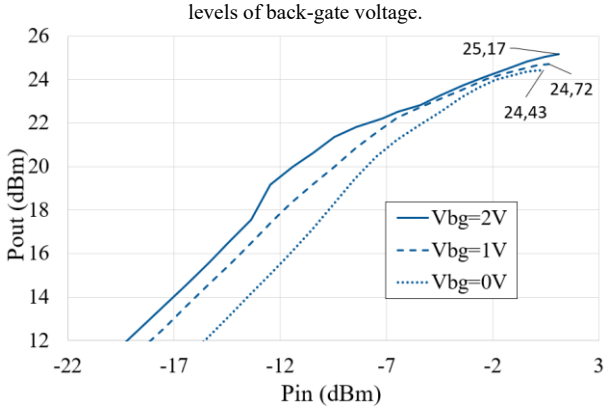


Figure 20. 28nm FDSOI PA Pout vs Pin measured performances for 3 levels of back-gate voltage.

The gain and PAE performances related to the frequency sweep from 1 GHz to 3 GHz for the 130 nm and 28 nm PA are presented in Figs. 17 and 18, respectively. The 130 nm PA exhibits a maximum Pout between 25 dBm and 30 dBm in the frequency range from 1.4 GHz to 2.5 GHz, with measurement results closely matching those of PLS. A maximum PAE between 37% and 40% is observed from 1.5 GHz to 2.4 GHz. The 28 nm PA presents a maximum Pout of 28.22 dBm and a maximum PAE of 38.44% at 1.85 GHz.

B. Fine Tuning Gain with Back Gate Transistor Bias in 28nm FDSOI Technology

In CMOS SOI technology, access to the transistor's back-gate provides additional control over the device's characteristics that can be leveraged to modify key performance parameters of a PA, such as output power, gain, and PAE. Changing the back-gate bias (*Vbg*) effectively modulates the transistor's threshold voltage *Vth*. A lower threshold voltage can increase the transistor's current driving capability, potentially enhancing the power output and, depending on the biasing conditions, the gain. However, this can also lead to higher power consumption and decreased efficiency.

The measured performance of gain versus P_{out} and P_{out} versus P_{in} for the PA in 28 nm with three different levels of *Vbg* voltage are presented in Figs. 19 and 20. For *Vbg=2V,* the transistors are more biased for maximum conduction, resulting in the highest initial gain of 31.3 dB and a curve with the typical shape of a class AB PA, reaching 21.8 dBm of linear output power. A *Vbg=1V* offers a more balanced operation, with a lower initial gain (30.2 dB) but greater linearity up to higher output power levels (*OCP1*=22.2 dBm). Meanwhile, *Vbg=0V* shows the lowest gain (27.7 dB) due to reduced transistor conduction, but the highest linear output power (*OCP1*=23.6 dBm). Despite the difference in maximum linear power output among the three bias setups, the maximum saturated power outputs are between 24.43 dBm and 25.17 dBm.

These results demonstrate how the back-gate voltage in 28nm FDSOI technology can be leveraged to optimize amplifier performance according to specific requirements for gain and linearity.

C. 130nm and 28nm Power Amplifier Comparative Analysis

This subsection presents a comparative analysis of the two PAs in terms of size and performance. As seen earlier at the start of the results section, the PA implemented in 28 nm technology occupies an area corresponding to 34% of the area occupied by the PA in 130 nm technology.

By integrating the balun functionality directly into the cascode input stage, the design removes the need for an external passive balun. This element typically occupies a significant portion of the RF front-end layout. Passive baluns rely on inductors and transformers with relatively large

TABLE I. COMPARING WITH THE STATE-OF-ART

Ref.	Freq. (GHz)	Psat (dBm)	P1dB (dBm)	PAE max (%)	PAE 6dB (%)	Gain (dB)	Topology	Technology	Supply (V)
[28]	2.3	32.8	32	59	40	27.5	LDMOS Doherty	130nm SOI**	3.4
[29]	2.4	35.1	34	53		29.5	Doherty	130nm SOI	5
[23]	1.95	30.5	29.7	53	40	26.5	Doherty	180nm SOI	4
[24]	1.85	31.9	N/A	56.2		14.2	ET PA	180nm bulk	4
[27]	2.6	33.1	N/A	43.5	N/A	28.1	4-stack E/Fodd	45nm SOI	3
[25]	2.4	30.3	N/A	36.5	29.1	N/A	C-commuted	40nm Bulk	2.4
[22]	2.4	31.6	N/A	49.2		N/A	Digital Outphasing	45nm bulk	2.4
[26]	1.85	30.7	28.8	44.4	28	11	Quasi-Doherty	180nm SOI	3
PA 130*	**1.85**	**32**	**30**	**49**	**26.6**	**34**	**Cascode Classe-AB**	**130nm PDSOI**	**5**
PA 28*	**1.85**	**28.8**	**28.3**	**38.5**	**20.8**	**33**	**Triple stack Classe-AB**	**28nm FDSOI**	**3**

*PLS | **SOI with LDMOS option

geometries, which scale poorly in deep-submicron CMOS. Replacing them with an active balun not only reduces the overall silicon footprint but also enables a more compact and area-efficient PA architecture, which is particularly advantageous for highly integrated IoT and multi-antenna systems.

A comparative analysis between Figs. 13 and 14 shows that the PA based on 130 nm PDSOI technology outperforms the 28 nm FDSOI in terms of S-parameter performance. The *S21* gain of the 130 nm PA remains around 35 dB in the central range (1.6 to 2.3 GHz), while the 28 nm PA reaches 30 dB only in the range between 1.5 and 1.9 GHz. However, the *S22* and *S11* of the 28 nm PA are more negative (below -5 dB and -15 dB, respectively), indicating better impedance matching at the input and output, with lower signal reflection.

A performance comparison between PAs gain (dB) and PAE (%) through Figs. 15 and 16, considering the post-layout simulation, shows that the 130 nm PA achieves higher maximum output power (~31 dBm) than the 28 nm PA (~28.5 dBm), making it more suitable for high-power applications. Considering the PLS performance, the 130 nm PA achieved a saturated output power of 31 dBm and the 28 nm PA achieved approximately 28.5 dBm, making the 130 nm technology more suitable for high-power applications, as expected. The 130 nm amplifier also provides slightly higher gain at lower output power levels. Furthermore, the 130 nm PA shows superior PAE performance in PLS, achieving a maximum of 48.5%, while the 28 nm PA achieves 38.35%. Comparing measurements, Fig. 15 shows that the 130 nm PA achieves a P_{sat} of 28.82 dBm and a P1dB of 27.29 dBm, while the 28 nm PA, in Fig. 16, reaches a P_{sat} of approximately 25.5 dBm and a P_{1dB} of 23.6 dBm.

Although the performance values of the circuit made with 130 nm technology are higher, the circuit in 28 nm technology allows for adjustments in gain and linearity performance through back-gate voltage. This enables the choice to operate in either a high-gain mode or a high-linearity mode, depending on the communication requirements.

D. State-of-the-Art Analysis

A comparison with the state of the art is conducted to conclude the performance assessment of the PAs presented in this section. Table I summarizes the state-of-the-art PAs and the performance metrics of the PAs developed in this research.

Considering 5G and NB-IoT applications that require modulations with high PAPR, the comparison was primarily made with promising topologies and techniques, such as Doherty and Envelope Tracking, as well as other high-efficiency classes.

It is observed that the two developed PAs outperform all PAs in Table I in terms of gain. Regarding *Psat*, the PA in 130 nm outperforms the works [22] [23] [24] [25] [26]. Regarding PAE, the PA in 130 nm outperforms the works [25] [26] [27], and the PA in 28 nm outperforms the work [25].

Regarding output power, the developed PAs are promising, as they are being compared with Doherty PAs, which consist of two or more PAs in parallel. If double the power were considered for the presented PAs, they would be comparable to Doherty's maximum power-level topologies.

The 130nm PDSOI pseudo-differential PA demonstrates an overall performance superior to the 28nm FDSOI design. PAs in [28] and [29] use off-chip passive components, which enhance performance due to significantly higher-quality factors than integrated passives. The PA architecture in [24] employs an envelope tracking technique, yielding a substantial improvement in PAE. Lastly, PA [25] is based on a switched amplifier architecture, enabling higher power density.

The PA designed in 28nm FDSOI is competitive in terms of state-of-the-art performance; however, the low quality factor of integrated passives tends to reduce the maximum achievable PAE.

V. CONCLUSION

This paper compared two PAs implemented in 130 nm PDSOI and 28 nm FDSOI technologies and targeted at low-power RF applications. The results section presented a detailed analysis of their S-parameter responses, gain, PAE, output power, frequency behavior, and, for the 28 nm PA, the impact of back-gate biasing.

The paper compares passive elements of both technologies through the resistivity of metal layers, capacitances, and inductances. It also compares active components (transistors), showing that the 130 nm PDSOI technology has much thicker layers than the 28 nm FDSOI technology, making it more suitable for power emission. However, the 28nm technology also enables this functionality while occupying three times less space, albeit at a considerably higher cost and with lower

performance, given its primary orientation towards digital circuits.

Both PAs are composed of a gain stage (driver) and a power stage (power amplifier, PA), utilizing differential and cascode topologies. The PA implemented in 28 nm technology features a 3-stacked transistor in its power stage, allowing for a higher drain bias voltage. This adjustment was deemed fair within the functional comparison, as the technology features thinner layers, necessitating such adaptations. Details about the interstage matching wideband transformer of both circuits were presented, and the cascode active balun circuit on the 28 nm PA was detailed.

Performance graphs were presented for S-parameters, PAE, gain, Pout, frequency sweep, and different biasing setups for the back-gate voltage in the 28 nm technology. The results indicate that the performance of the circuit fabricated in 130 nm technology is superior to that of the 28 nm circuit.

The 28 nm FDSOI technology enables fine-tuning of the PA's gain through back-gate voltage, thus providing additional operational freedom. The two employed technologies, 130 nm and 28 nm, can produce PAs suited for the intended application.

The developed PAs exhibit superior gain performance compared to the state-of-the-art. They are promising in power when used in efficiency-boosting topologies that combine multiple PAs to increase PAE at backoff and maximize output power.

For future research, it is suggested that we explore the use of these PAs in efficiency-enhancing topologies and power-combining strategies, such as Doherty and Envelope Tracking, to facilitate comparisons with the state-of-the-art and contribute to the development of circuits for 5G and NB-IoT applications.

ACKNOWLEDGMENT

This study was financed in part by the project Beyond5 and by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil* (CAPES) - Finance Code 001.

REFERENCES

[1] M. L. Carneiro, T. LeCocq, E. Kerherve, M. Matos, T. Taris and J. Pham, *Comparative Analysis of 130nm PDSOI and 28nm FDSOI Technologies for 5G Power Amplifier Applications,* The Twenty-First International Conference on Wireless and Mobile Communications (ICWMC 2025) - IARIA, v.21, pp. 10 - 15, Lisbon, 2025.

[2] K. Mekki, E. Bajic, F. Chaxel and F. Meyer, *Overview of Cellular LPWAN Technologies for IoT Deployment: Sigfox, LoRaWAN, and NB-IoT,* IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), pp. 197–202. doi: 10.1109/PERCOMW.2018.8480255, 2018.

[3] G. K. Celler and S. Cristoloveanu, *Frontiers of Silicon-on-Insulator,* Journal of Applied Physics, vol. 93, pp. 4955–4978, 2003.

[4] H. Mendez, *Silicon-on-Insulator, SOI Technology and Ecosystem, Emerging SOI Applications,* April, 2009.

[5] A. Chen, *Advances in Semiconductor Technologies: Selected Topics Beyond Conventional CMOS,* Wiley-IEEE Press, 2023.

[6] A. S. Gillis, *What is the Internet of Things (IoT)?,* Available: https://www.techtarget.com/iotagenda/definition/Internet-of-Things-IoT. [Accessed: Mar. 28, 2023], 2014.

[7] Telefonaktiebolaget LM Ericsson, *What is 5G? How will it transform our world?,* Available: https://www.ericsson.com/fr/5g. [Accessed: Mar. 30, 2023], 2021.

[8] P. Wegner, *Global IoT market size to grow 19% in 2023 - IoT shows resilience despite economic downturn,* Available: https://iot-analytics.com/iot-market-size. [Accessed: Mar. 28, 2023], 2023.

[9] Thales Group, *Massive IoT: Tech overview, business opportunities and examples,* Available: https://www.thalesgroup.com/en/markets/digital-identity-and-security/mobile/massive-iot. [Accessed: Mar. 30, 2023], 2022.

[10] GSMA Mobile IoT Industry Alignment group, *3GPP Low Power Wide Area Technologies,* Svetlana Grant (GSMA), September, 2016.

[11] A. T. Prittu and M. Mathurakani, *SC-FDMA -An Efficient Technique For PAPR Reduction In Uplink Communication Systems -A Survey,* IJRET: International Journal of Research in Engineering and Technology, Volume: 03 Special Issue: 01 | NC-WiCOMET-2014, 2014.

[12] ETSI, *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) conformance specification; Radio transmission and reception; Part 1: Conformance testing (3GPP TS 36.521-1 version 16.8.1 Release 16),* Available: https://www.etsi.org/deliver/etsi_ts/136500_136599/13652101/16.08.01_60/ts_13652101v160801p.pdf. [Accessed: Mar. 31, 2023], 2021.

[13] GSM Association, *NB-IoT Deployment Guide to Basic Feature set Requirements,* June, 2019.

[14] Ixia, *SC-FDMA, Single Carrier FDMA in LTE,* 915-2725-01 Rev A, November 2009.

[15] H. G. Myung, J. Lim and D. J. Goodman, *Peak-To-Average Power Ratio of Single Carrier FDMA Signals with Pulse Shaping,* IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 1–5. doi: 10.1109/PIMRC.2006.254407, 2006.

[16] N. Alfaraj, *SIMOX vs. Smart-Cut: A Brief Overview of the Silicon-on-Insulator Technology,* doi: 10.13140/RG.2.2.10340.27522, January, 2018.

[17] Renesas Electronics, *Semiconductor Reliability Handbook,* R51ZZ0001EJ0250 Rev. 2.50, January, 2017.

[18] A. D. Pham, *Biasing Techniques for Linear Power Amplifiers,* Ph.D. dissertation, Massachusetts Institute of Technology, Available: https://api.semanticscholar.org/CorpusID:112109794, May, 2002.

[19] IBM Corporation, *BEOL Process Challenges - Emerging CMOS Technology at 5 nm and Beyond,* IEDM 2015 Short Course, Albany Nano Tech Center, NY, USA, 2015.

[20] B. Leite, *Design and modeling of mm-wave integrated transformers in CMOS and BiCMOS technologies,* University of Bordeaux, 2011BOR14359, PhD Thesis, address: http://www.theses.fr/2011BOR14359/document, 2011.

[21] Y. Sim, J. Park, J. Yoo, C. Lee and C. Park, *A CMOS power amplifier using an active balun as a driver stage to enhance its gain,* Microelectronics Journal, vol. 63, 05, 2017.

[22] A. Banerjee, L. Ding and R. Hezar, *A High Efficiency Multi-Mode Outphasing RF Power Amplifier With 31.6 dBm Peak Output Power in 45nm CMOS,* IEEE Transactions on Circuits and Systems I : Regular Papers, vol. 67, no 3, pp. 815-828, doi: 10.1109/TCSI.2019.2954068, 2020.

[23] P. Draxler and J. Hur, *A Multi-Band CMOS Doherty PA with Tunable Matching Network,* IEEE MTT-S International Microwave Symposium (IMS), pp. 944–946. doi: 10.1109/MWSYM.2017.8058742, 2017.

[24] S. Jin, B. Park and K. Moon, *A Highly Efficient CMOS Envelope Tracking Power Amplifier Using All Bias Node Controls,* IEEE Microwave and Wireless Components Letters, vol. 25, no. 8, pp. 517–519, 2015.

[25] B. Yang, H. J. Qian and X. Luo, *26.5 A Watt-Level Quadrature Switched/Floated Capacitor Power Amplifier with Back-Off Efficiency Enhancement in Complex Domain Using Reconfigurable Self-Coupling Canceling Transformer,* IEEE International Solid-State Circuits Conference (ISSCC), pp. 362-364. doi :10.1109/ISSCC42613.2021.9366029, 2021.

[26] K. Kim, D.-H. Lee and S. Hong, *A Quasi-Doherty SOI CMOS Power Amplifier With Folded Combining Transformer,* IEEE Transactions on Microwave Theory and Techniques, vol. 64, no 8, pp. 2605-2614, doi : 10.1109/TMTT.2016.2577584, 2016.

[27] M. Khorshidian and H. Krishnaswamy, *A Fully-Integrated 2.6GHz Stacked Switching Power Amplifier in 45nm SOI CMOS with >2W Output Power and 43.5% Efficiency,* IEEE MTT-S International Microwave Symposium (IMS), pp. 323–326, 2019.

[28] A. Serhan, D. Parat and S. Gerardin, *A Reconfigurable SOI CMOS Doherty Power Amplifier Module for Broadband LTE High-Power User Equipment Applications,* IEEE Radio Frequency Integrated Circuits Symposium (RFIC), pp. 79–82. doi: 10.1109/RFIC49505.2020.9218305, 2020.

[29] P. Reynier, A. Serhan, D. Parat and S. Gerardin, *A High-Power SOI-CMOS PA Module with Fan-Out Wafer-Level Packaging for 2.4 GHz Wi-Fi 6 Applications,* IEEE Radio Frequency Integrated Circuits Symposium (RFIC), pp. 59–62, 2021.

[30] T. Shen, W. Zhang, K. B. Yeap, J. Tan and P. Justison, *An Investigation of Dielectric Thickness Scaling on BEOL TDDB,* IEEE International Reliability Physics Symposium, pp. 3A.2.1–3A.2.6. doi: 10.1109/IRPS.2015.7112698, 2015.

[31] Y. Eo and K. Lee, *High Efficiency 5GHz CMOS Power Amplifier with Adaptive Bias Control Circuit,* IEEE Radio Frequency Integrated Circuits (RFIC) Systems, Digest of Papers, pp. 575–578. doi: 10.1109/RFIC.2004.1320686, 2004.

Leveraging Asset Administration Shells and Fog Computing for Scalable and Secure Smart Pool Management

André C. Costa[1], Rui Pinto[1,2], Gil Gonçalves[1,2]
Dept. of Informatics Engineering, Faculty of Engineering, University of Porto, Porto, Portugal[1]
SYSTEC, ARISE, Faculdade de Engenharia, Universidade do Porto, Porto, Portugal[2]
Email: up201905916@up.pt, {rpinto, gil}@fe.up.pt

***Abstract*—Maintaining optimal water quality in swimming pools is critical for ensuring safety, hygiene, and user comfort, yet traditional methods often prove time-consuming, error-prone, and operationally inefficient. This paper presents SmartPool, an advanced Cyber-Physical System (CPS)-based solution that automates swimming pool management through real-time monitoring, automated control, and Asset Administration Shell (AAS)-based Digital Twins. Unlike existing proprietary ecosystems, SmartPool leverages AAS as a standardized framework to ensure interoperability, scalability, and lifecycle-oriented management of diverse pool assets, moving beyond closed solutions. The system is built upon a robust five-level CPS architecture, strategically incorporating a Fog computing layer for distributed intelligence, hosting the AAS environment, and facilitating efficient edge processing. SmartPool measures critical water parameters such as pH, chloride, temperature, and water levels, provides real-time data visualization through a user-friendly dashboard, and enhances safety with integrated camera and recognition algorithms. This extended work details the comprehensive methodology, the enhanced architectural design, and the prototype validation, demonstrating the feasibility of this open, AAS-enabled approach to significantly improve efficiency, sustainability, and safety in pool operations. The research lays the foundation for scalable smart water management systems and promotes seamless integration into broader Industry 4.0 and smart city initiatives.**

Keywords-Pool System; Pool Maintenance; Automation; Control; Asset Administration Shell

I. Introduction

This work builds upon a preliminary conference contribution that introduced the *SmartPool* concept as an automated Cyber-Physical System (CPS) for swimming pool monitoring and safety [1]. That initial study demonstrated the feasibility of integrating sensors, actuators, and real-time monitoring into a unified system for pool management. The present article extends that work by significantly deepening the architectural design, expanding the prototype implementation, and providing a more comprehensive validation and discussion of results.

Swimming pools are an increasingly common feature in modern environments, serving recreational, fitness, hospitality, and tourism purposes. With over 10.7 million public and residential pools in the United States alone [2], the global expansion of aquatic facilities places growing pressure on pool maintenance systems to meet expectations for safety, sustainability, and operational efficiency [3]. These demands are further intensified by frequent water consumption and chemical treatments, which impact both environmental sustainability and user well-being.

Ensuring water quality and safety is paramount. Inadequate pool maintenance is directly associated with health risks, including dermatological conditions, respiratory irritants, and waterborne disease outbreaks such as cryptosporidiosis and giardiasis [4]. These challenges highlight the need for intelligent, accountable, and data-driven pool management systems that go beyond isolated sensing and manual intervention.

While existing market solutions already incorporate automated control logic—typically through sensors and chemical dosing actuators—most offerings remain proprietary, closed, and vertically integrated. Such architectures limit interoperability, hinder data exchange with external platforms, and provide limited support for asset lifecycle visibility. Moreover, the fragmentation of sensor data, chemical consumption records, and equipment diagnostics prevents holistic oversight, particularly in multi-stakeholder or public pool environments.

Recent advances in CPS, the Internet of Things (IoT), and Digital Twin (DT) technologies [5], [6] offer a promising path forward. IoT enables continuous data acquisition from distributed sensors, CPS supports tight integration between physical and digital layers, and DTs facilitate simulation, forecasting, and lifecycle-oriented management. Among DT frameworks, the Asset Administration Shell (AAS) [7] has emerged as a standardized and semantically rich representation capable of unifying heterogeneous devices into interoperable systems [8]–[10].

In this context, this paper presents an extended version of *SmartPool*, a CPS-enabled prototype platform that applies AAS-based Digital Twins to swimming pool environments. The main goals of this work are to:

- Design an open, modular architecture for intelligent pool monitoring and control, supporting integration with IoT sensors and visual safety systems.
- Apply AAS-based DTs to pool assets—such as sensors, actuators, and tanks—to provide structured, interoperable digital representations.
- Develop and validate a working prototype that demonstrates real-time monitoring, automated actuation, and object recognition-based safety mechanisms.
- Explore the feasibility of lifecycle-oriented, standards-driven asset management in non-industrial water environments.

Unlike existing proprietary solutions, *SmartPool* leverages open-source middleware (*Eclipse BaSyx* [11], [12]) and a CPS-inspired architecture to support bi-directional communication

between physical components and digital services [13]. Visual surveillance and object recognition enable proactive safety mechanisms, while AAS *Submodels* expose asset states and metadata, supporting transparency, traceability, and integration with smart building and smart city ecosystems.

Developed as an academic proof-of-concept under resource constraints, the system demonstrates the practical applicability of CPS and AAS principles to domains traditionally excluded from industrial digitalization. The research contributes to ongoing discussions on intelligent water management [14] by introducing a standards-aligned, modular framework that can be extended to public facilities, hospitality venues, and cyber-physical smart environments.

The remainder of this paper is structured as follows. Section II reviews relevant literature on automated pool systems and digitalization approaches. Section III presents the adopted research methodology. Section IV details the system architecture and components. Section V describes the prototype implementation and validation. Section VI discusses the results, limitations, and implications for real-world deployment. Finally, Section VII concludes the paper and outlines future research directions.

II. Related Work

Beyond dedicated smart pool systems, the principles of intelligent monitoring and control are increasingly vital across various water control and maintenance domains, including urban water networks [15], industrial wastewater treatment [16], and agricultural irrigation [17]. These broader "smart water management systems" share common challenges in efficiency, sustainability, and data-driven decision-making, which IoT, CPS, and DTs are actively addressing [18].

In these diverse water management areas, IoT technologies enable continuous sensing and data acquisition from a myriad of distributed points. For instance, urban water networks [15] deploy sensors to monitor water flow, pressure, leakage, and quality parameters like turbidity and chlorine residuals, ensuring efficient distribution and early detection of infrastructure issues. Industrial wastewater facilities [16] use sensors to track chemical composition, temperature, and pH levels to comply with environmental regulations and optimize treatment processes. Similarly, in agriculture [17], soil moisture, nutrient levels, and weather data are collected to inform precision irrigation, minimizing water waste. These systems often utilize various sensors for pH, temperature, turbidity, water level, and water flow.

CPS architectures [19] are employed to create closed feedback loops between the physical environment and digital control, enabling automated adjustments. For example, in smart irrigation, sensor data automatically triggers actuators (like valves or pumps) to deliver water precisely where and when needed. In industrial settings, real-time water quality data can activate dosing pumps or filtration systems to maintain desired parameters. The communication backbone frequently relies on protocols like MQTT [20] for efficient, real-time data exchange from edge devices. For specialized applications, such as underwater monitoring in reservoirs or rivers, technologies like IoT-LoRa [21] are explored, though they face challenges with water type and turbidity.

Furthermore, DTs provide advanced monitoring, predictive analysis, and lifecycle management capabilities in these sectors. By creating virtual replicas of physical water infrastructure—be it a section of a city's pipeline, a specific industrial treatment unit, or an agricultural field—operators can simulate scenarios, predict maintenance needs, and optimize resource allocation. These systems often leverage Edge and Fog computing paradigms [22], with devices like Raspberry Pis and Arduinos [23] serving as aggregation nodes or local processing hubs, offloading computational responsibilities from central cloud servers and enabling faster response times for critical events. The processed data can then be sent to the cloud for extra storage and computational analysis.

A. Literature Review

Modern swimming pool management has already reached a considerable degree of automation [24]–[27]. In practice, machine rooms are typically equipped with sampling points, where sensors continuously measure water quality parameters such as pH, chlorine concentration, and turbidity. These readings are fed into a controller with sufficient computational power to analyze the data and activate actuators that dose chemicals or adjust filtration cycles to maintain water quality within predefined setpoints. Major vendors also provide mobile applications and integrated ecosystems that enable companies and pool operators to remotely monitor and adjust pool parameters. However, these systems are proprietary and vertically integrated, often limiting interoperability, extensibility, and long-term integration with broader smart environments.

Academic research, on the other hand, has explored swimming pool management mostly through prototypes and small-scale experiments. These studies often emphasize either water quality monitoring or safety/drowning prevention, presenting diverse architectures and sensing approaches. To provide a structured overview, Table I summarizes the main contributions of relevant works across these domains. The following subsections briefly review each research stream.

a) ***Water Quality Monitoring****:* Monitoring water quality [34] is a fundamental requirement for ensuring hygiene, comfort, and safety in swimming pools. Parameters such as pH, turbidity, temperature, and water level must be continuously monitored to maintain optimal conditions and meet health regulations. Recent works have explored automated, sensor-based systems for real-time tracking of these metrics. For instance, Hamid *et al.* [28] proposed a Smart Water Quality Monitoring System (SWQMS) capable of tracking pH and temperature in real time. Their findings confirmed that while temperature fluctuates throughout the day, pH levels remain relatively stable, underscoring the need for automated monitoring. However, their system was limited in scope, tracking only a narrow set of parameters and lacking integration with broader pool management functions. Expanding on this, Lakshmikantha *et al.* [29] introduced a more comprehensive

TABLE I. Summary of related works in swimming pool monitoring and management.

Author / Year	Focus Area	Parameters / Sensors	Architecture / Technology	Limitations
Hamid *et al.* [28]	Water quality monitoring	pH, temperature	SWQMS prototype, automated tracking	Limited scope (only two parameters); lacks explicit standardization (AAS) for interoperability or a comprehensive multi-level CPS architecture
Lakshmikantha *et al.* [29]	Water quality monitoring	pH, turbidity, temperature, level, flow	Multi-sensor IoT prototype	Prototype stage, limited validation; does not address standardized asset representation (AAS) or a comprehensive multi-level CPS architecture for lifecycle management
Sangeetha *et al.* [30]	Safety / drowning prevention	Ultrasonic, PIR	SSPMS with alarms and drainage	Safety-focused, no water quality integration; proprietary approach (implied), lacks standardized interoperability (AAS) or a holistic multi-level CPS framework
Raj *et al.* [31]	Safety + monitoring	Temperature, level, intoxication detection	ESP32-CAM with IoT connectivity	Limited scalability, prototype; primarily focused on safety with specific hardware, no explicit mention of open standards like AAS for broader interoperability or a comprehensive multi-level CPS architecture
Christopher *et al.* [32]	Communication challenges	LoRa signals underwater	IoT-LoRa tests in various water types	Performance degradation in seawater; focuses on communication layer, does not address overall system architecture, standardized asset representation (AAS), or interoperability with other systems
Glória *et al.* [33]	IoT architectures	Multiple environmental sensors	Raspberry Pi + Arduino via MQTT	Generic, not pool-specific; while exploring IoT gateways, it does not explicitly integrate standardized digital twins (AAS) or a multi-level CPS architecture for complex asset management and interoperability across varied smart environments

prototype that included monitoring of turbidity, water level, and flow. While their solution was more versatile and adaptable to other water contexts such as industrial wastewater, it remained largely a prototype, with no deployment-focused validation, energy optimization, or discussion on connectivity and security challenges.

b) ***Safety and Drowning Prevention****:* Safety and drowning prevention [35] are crucial components of intelligent pool systems, especially given the high rates of drowning-related incidents, particularly among children. To address this, researchers have developed systems that fuse environmental sensing and visual analysis to detect emergency situations in real time. Sangeetha *et al.* [30] presented a Smart Swimming Pool Management System (SSPMS) combining ultrasonic and Passive Infrared (PIR) sensors for real-time drowning detection and emergency response. Although promising, their system was reactive rather than predictive and relied on basic threshold-based logic, limiting robustness in complex scenarios. Raj *et al.* [31] advanced this concept by integrating ESP32-CAM-based [36] image analysis to identify intoxicated individuals or drowning victims, alongside water level and temperature sensing. Their system also featured emergency notifications. However, this approach depends heavily on stable lighting conditions and the accuracy of basic computer vision models, which can be unreliable in dynamic pool environments. Furthermore, privacy concerns associated with visual surveillance were not thoroughly addressed, nor was the system evaluated under real-world operating conditions.

c) ***Communication and Connectivity****:* Effective communication is critical in distributed sensing systems, especially in aquatic environments where signal propagation can be degraded by factors like water salinity, turbidity, and interference [37]. The use of long-range, low-power protocols such as LoRaWAN and MQTT [38] is gaining traction due to their energy efficiency and robustness in constrained environments. Christopher *et al.* [32] investigated IoT-LoRa [21] signal behavior in water environments and found that pool water, due to its lower salinity, supports better signal transmission compared to seawater. Their findings help guide protocol selection based on fluid properties. However, their analysis focused primarily on static conditions and lacked insights into performance under user-generated interference (e.g., people swimming), device mobility, or real-time data throughput. Additionally, their study did not explore multi-hop or hybrid communication models, which are often necessary in complex installations.

d) ***Architectural Innovations****:* Modern smart pool systems increasingly benefit from hybrid architectures that leverage the complementary strengths of Edge and Cloud computing [39]. These enable real-time decision-making at the Edge while facilitating advanced analytics and long-term storage in the Cloud. Glória *et al.* [33] demonstrated such a layered architecture using a Raspberry Pi as an IoT gateway and Arduino-based sensor nodes for data acquisition. Data communication was handled via the MQTT protocol [20], providing a lightweight and reliable channel for telemetry. While this architecture offers flexibility and modularity, it lacks built-in security mechanisms, which are critical when dealing with safety-critical applications. Moreover, their work focused primarily on system architecture without delving into scalability, fault tolerance, or orchestration aspects. Later work by Andriulo *et al.* [22] explored more sophisticated Cloud–Edge integration, but the computational burden on local devices remained a constraint, especially for tasks involving visual processing or real-time inference.

B. Summary and Gap

The reviewed literature highlights important contributions to monitoring and safety in swimming pools. However, most

academic studies adopt an experimental or prototype perspective, often neglecting the fact that commercial pool systems already provide closed-loop automation in practice. Moreover, the academic focus tends to shift toward safety mechanisms (especially drowning prevention) rather than continuous, standardized, and interoperable water quality management [40]. This creates a gap between industrial practice and academic research. While industry offers proprietary ecosystems, research opportunities remain in the application of open, standardized approaches that support lifecycle management of all pool assets—including sensors, actuators, water, and chemicals.

A significant limitation observed across many of these smart water solutions, akin to commercial smart pool systems, is their tendency to be proprietary, closed, and vertically integrated. This often limits interoperability, data sharing, and seamless integration with broader smart environments or third-party services. This fragmentation hinders comprehensive lifecycle management of heterogeneous assets and impedes the realization of truly interconnected ecosystems.

This is where the AAS concept, central to the *SmartPool* solution, offers a transformative approach. As a standardized implementation of Digital Twins, AAS represents a mechanism to unify heterogeneous devices and data under a common data model, ensuring interoperability, scalability, and long-term maintainability. In urban water management, each pump, valve, or sensor could have its own AAS, allowing different manufacturers' equipment to communicate and be managed uniformly. In industrial wastewater, each chemical dosing unit, filter, or monitoring probe could expose its functions and data through a standardized AAS, facilitating integration with factory-wide Industry 4.0 systems [41]. The *SmartPool* system, by leveraging AAS to digitalize pool assets and enable bi-directional communication with standardized interfaces, demonstrates how open and interoperable architectures can enhance transparency, scalability, and integration with broader Industry 4.0 and smart city initiatives, effectively laying the foundation for more interoperable smart water systems.

In this regard, the present work leverages the concept of AAS as the standardized digital representation of assets. By adopting an AAS-based architecture implemented through *Eclipse BaSyx* [11], [12], *SmartPool* demonstrates how pool automation can move beyond proprietary solutions toward interoperable, research-driven, and lifecycle-oriented management.

III. Methodology

The development of the *SmartPool* system followed a structured Design Science Research (DSR) methodology [42], widely used in CPS and IoT engineering contexts. This methodology was organized into three iterative and interdependent phases: (1) problem analysis and requirement elicitation, (2) system design and technology selection, and (3) implementation and prototype validation. Each phase was informed by engineering best practices, existing standards, and empirical feedback from iterative development cycles.

A. Requirements Elicitation and Problem Definition

The first phase focused on systematically identifying the functional and non-functional requirements for a smart pool management system. Functional requirements were derived from the domain-specific needs of aquatic facilities and centered on the following objectives:

- Continuous sensing of water quality parameters (e.g., pH, temperature, turbidity);
- Automated actuation mechanisms for water circulation and dosing;
- Real-time monitoring and alerts for safety-critical events (e.g., falls or drowning risks);
- Integration of a unified user interface for administrators and end-users.

Non-functional requirements were equally considered, emphasizing:

- Interoperability, to allow seamless integration of heterogeneous sensors and actuators;
- Scalability, to support deployment across pools of varying size and complexity;
- DT compatibility, through the use of standardized digital representations of assets;
- Security and dependability, ensuring reliable and confidential communication across components.

These requirements were elicited through a combination of domain analysis, literature review, and benchmarking against existing smart pool systems.

B. Architectural Design and Technology Selection

The second phase applied architecture-driven engineering to explore and define the technical backbone of the *SmartPool* system. A comparative analysis of sensors, microcontrollers, communication protocols, and middleware platforms was conducted. Technologies were evaluated against a matrix of selection criteria, including compatibility with open standards, cost-effectiveness, modularity, and support for Industry 4.0 paradigms, particularly DTs and edge-cloud integration.

This process led to the adoption of a CPS-inspired architecture with the following characteristics:

- Edge Layer: Microcontroller-based sensor nodes and actuators responsible for local control and data acquisition;
- Fog Layer: A local processing hub (Raspberry Pi) acting as an MQTT broker and pre-processing node;
- Cloud Layer: Centralized services for data persistence, advanced analytics, and integration with user interfaces;
- DT Layer: Adoption of the AAS concept through the *Eclipse BaSyx* middleware [11], [12], enabling the standardized digital representation of physical pool components.

Communication protocols were selected to meet the dual goals of lightweight messaging and secure interoperability. MQTT over Transport Layer Security (TLS) encryption [43] was chosen for sensor-to-fog and fog-to-cloud communication, ensuring encrypted, publish/subscribe-based messaging with low latency.

C. System Implementation and Prototype Validation

In the final phase, a fully functional proof-of-concept prototype was implemented to validate the proposed architecture and assess system behavior under real-world constraints. The prototype integrated:

- Real-time water quality monitoring using calibrated sensors;
- Automated actuation based on rule-based logic (e.g., activating filters or chemical dispensers);
- Live video processing for human presence detection;
- Secure data transmission across all CPS layers;
- AAS-based digital twins for each key pool component, enabling dynamic querying and status tracking.

Validation involved both qualitative and quantitative evaluation, including:

- System responsiveness: Time-to-alert and actuation latency were measured for safety events and water quality thresholds.
- Communication performance: Throughput and reliability of MQTT communication under typical network loads.
- DT operations: Correctness of lifecycle events and API interactions via the AAS interface.
- User feedback: Functional testing of the User Interface (UI) with simulated users to assess usability and visualization of sensor data.

These validation efforts demonstrated the feasibility and coherence of the architecture, and also helped identify areas for future optimization, particularly in terms of energy efficiency and Artificial Intelligence (AI)-based analytics.

IV. SmartPool Solution

SmartPool was conceived as a research-driven platform that combines existing principles of pool automation with open-source frameworks and standardized DT representations. The system integrates multi-parameter sensing, real-time actuation, and lifecycle-oriented digitalization through the AAS. In doing so, *SmartPool* moves beyond the proprietary solutions offered by industry vendors, demonstrating how open and interoperable architectures can enhance transparency, scalability, and integration with broader Industry 4.0 ecosystems.

The architecture follows a five-level CPS-inspired structure [19] (Figure 1), which organizes perception, communication, middleware, application, and business functions. This design ensures modularity and separation of concerns, while supporting interoperability across heterogeneous devices and services.

a) ***Perception Level****:* The physical layer integrates all assets responsible for data acquisition and actuation. Water quality sensors measure parameters such as turbidity, pH, chlorine concentration, and temperature, while environmental sensors capture auxiliary data (e.g., light intensity). A camera provides real-time video streams for safety monitoring. Actuators regulate chemical dosing, lighting, and alarms, enabling both automatic control and safety notifications.

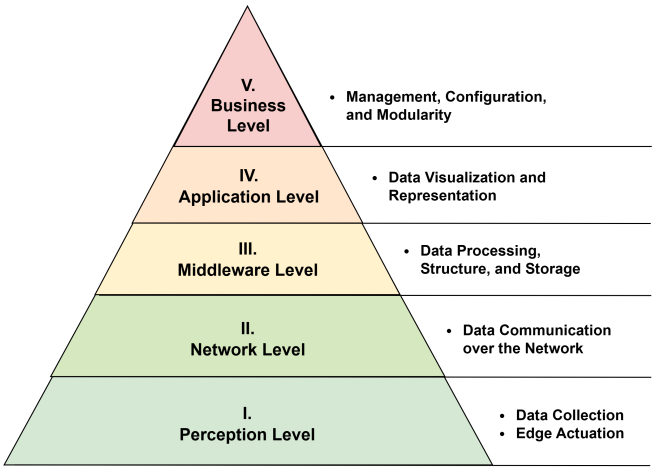


Figure 1. 5-Level Architecture

b) ***Network Level****:* This level manages connectivity between devices and higher layers. A central processing unit aggregates sensor data locally and coordinates actuation. Connectivity is supported by Wi-Fi [44] and communication protocols such as MQTT [45], OPC UA [46], [47], or Modbus TCP [48], ensuring compatibility with heterogeneous devices and infrastructures.

c) ***Middleware Level****:* Acting as the system's digital backbone, the middleware manages data processing, aggregation, and digitalization. The controller connects to a Fog node hosting **(i)** a message broker as intermediary for real-time data exchange, **(ii)** an AAS-based environment for standardized DT representations, and **(iii)** a time-series database for persistent storage. Each asset—including sensors, actuators, pool water, and dosing chemicals—is represented as an AAS instance, enabling lifecycle-oriented management.

Imagine every device in a pool – from the pH sensor to the filtration pump – doesn't just exist, but also carries a digital passport (its AAS). This passport isn't just a static document; it's a living, machine-readable record that contains all its essential information: its 'name' (identifier), technical specifications, real-time 'health' readings (sensor data), operational 'permissions' (actuator commands), and even its maintenance history.

Just like a country's border control system can quickly process individuals by scanning their standardized passports, the *SmartPool* system, using AAS, can seamlessly identify, monitor, and control any pool device, regardless of its manufacturer or specific communication protocol. Scaling 'SmartPool-as-a-Service' for multiple clients, it's like setting up a centralized agency that can issue, track, and manage these digital passports for thousands of pools. Each pool's collection of device passports ensures a uniform, secure, and interoperable way to manage everything, allowing for efficient, data-driven services on a large scale.

Through the AAS infrastructure, assets can be dynamically registered, monitored, and managed via standardized reposi-

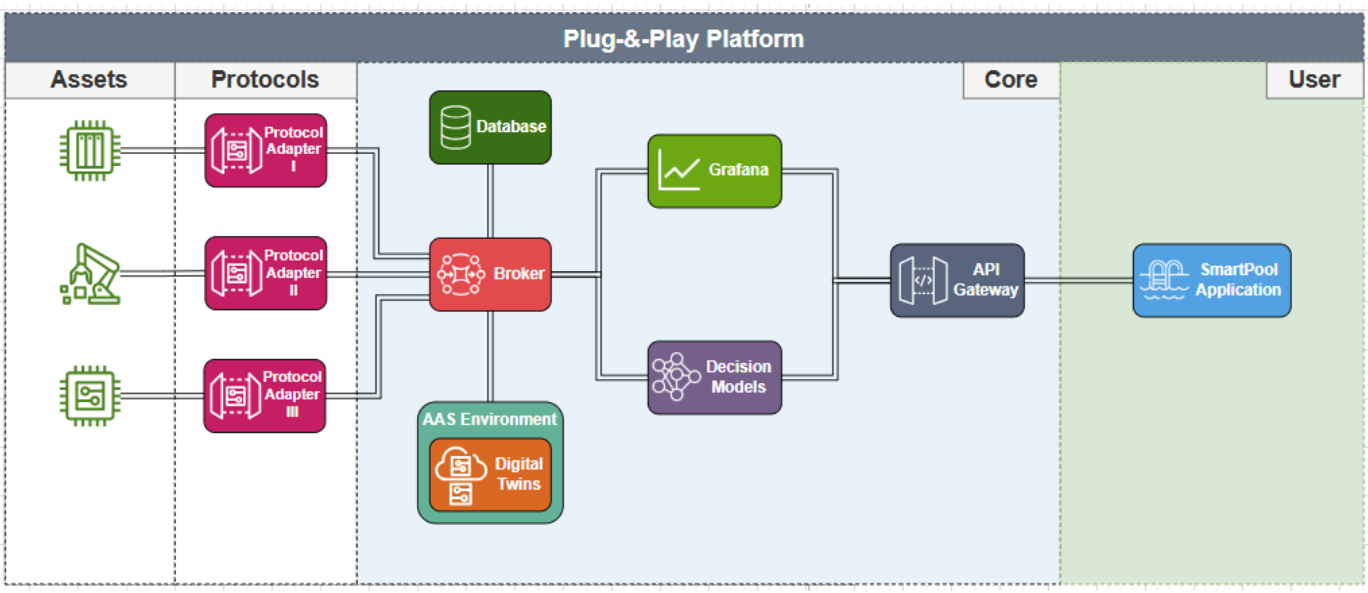


Figure 2. *SmartPool* System's Architecture.

tories and registries. The specification and representation of assets through AAS require several key considerations:

- **AAS Types and Instances:** The AAS framework distinguishes between *types* and *instances*. A type defines a general blueprint for an asset class, specifying its structure, mandatory *Submodels*, and semantic annotations. An instance, by contrast, represents a concrete, real-world manifestation of that asset, adhering to the schema defined by its corresponding type. Additionally, AAS types and instances are represented in an hierarchy deriving characteristics from parent representations [49].
- **Submodel Templates:** *Submodels* are collections of elements of information related to a specific aspect of an asset (e.g., technical data, digital nameplate, or lifecycle information) [49]. To ensure reusability and consistency, *Submodel Templates* are being defined and standardized by organizations such as ZVEI [50], the Industrial Digital Twin Association (IDTA) [51], and Plattform Industrie 4.0 [52]. These templates provide agreed structures for information exchange across domains, ensuring that assets of the same type can be represented in a consistent and machine-readable manner [53].
- **Semantic Descriptions:** To guarantee unambiguous interpretation of data, AAS elements must be semantically annotated. Semantic identifiers can be derived from established ontologies, data dictionaries, or classification systems such as IEC Common Data Dictionary (IEC CDD) and ECLASS [54], [55]. This semantic enrichment ensures that information is interoperable across systems and domains, supporting meaningful integration into wider ecosystems.
- **Interoperability:** Beyond digitalization, the AAS enables interoperability by providing standardized interfaces and semantics for asset communication. Achieving this requires collecting and harmonizing data from heterogeneous sources (e.g., sensors, controllers, and databases) into the AAS structure. Several open-source frameworks, including *Eclipse BaSyx* [56], CoreAAS, and PyI40AAS [57], actively support this integration by offering tools for modeling, registering, and exchanging AAS data.

d) ***Application Level****:* At this level, the *SmartPool* application provides real-time monitoring, historical dashboards, alerts, and control options. The application is designed to scale with the number of assets integrated into the system, supporting both individual pool owners and professional operators managing multiple facilities. Unlike commercial vendor-specific platforms, *SmartPool* emphasizes openness, allowing integration of safety features such as drowning detection (Section II) alongside water quality monitoring.

e) ***Business Level****:* The top layer focuses on decision support and lifecycle optimization. By leveraging AAS-based digital representations, *SmartPool* enables advanced analytics such as predictive maintenance, optimization of chemical usage, and integration with larger infrastructures (e.g., smart buildings or city-wide water management systems). This level transforms real-time monitoring into strategic insights for efficient, sustainable operation.

The components described across the five levels are illustrated in Figure 2, which depicts the *SmartPool* architecture and the interactions between its physical assets, middleware, and digital representations. The figure highlights how heterogeneous devices communicate through standardized interfaces, ultimately enabling interoperability, real-time monitoring, and lifecycle-oriented management of the pool system.

V. Prototype Validation

This section presents the proof-of-concept implementation of the *SmartPool* system introduced in the previous section.

The validation is structured into three parts: (i) monitoring and control of field assets, (ii) communication and digitalization of data into AAS-based DTs, and (iii) the *SmartPool* web environment created during implementation.

A. Monitoring and Control

As discussed in Section I, the prototype was constrained by limited hardware resources and the absence of laboratory conditions to fully replicate a pool environment. To represent the pool, a transparent plastic container was used, as shown in Figure 3.

The sensing setup included a temperature and humidity sensor (not waterproof, therefore positioned near the water surface), a light intensity sensor, and an ultrasonic sensor mounted at the top of the container to measure water level. The water level was calculated as the difference between the container height and the distance reported by the sensor.

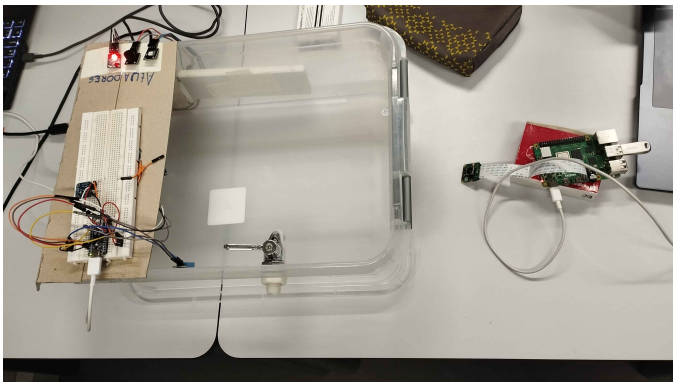


Figure 3. *SmartPool* Prototype Setup.

Actuation was represented by LEDs, which provided visual feedback when sensor readings exceeded predefined thresholds (e.g., red for low water level), and illuminated the container under low-light conditions. A buzzer simulated a safety alarm, alerting the user to the presence of entities such as a person or a pet near the water.

These devices were connected to microcontrollers responsible for data collection and actuation. Communication with the central node was established over Wi-Fi [44], using the MQTT protocol [45] to publish and subscribe data streams. Topics were organized hierarchically (e.g., `sensor/#` and `actuator/#`), with specific identifiers for parameters such as luminosity, distance, or temperature. For reliability, Arduinos [58] were configured with fault-tolerance mechanisms to attempt reconnection in case of Wi-Fi failures.

A Raspberry Pi acted as the primary processing and communication hub, equipped with a Raspberry Pi Camera Module 2 (Sony IMX219, 8 MP) [59] positioned above the container. In addition to capturing video, the Raspberry Pi executed object recognition using OpenCV [60] and the YOLOv8 model [61], detecting entities in the pool area and triggering alerts, thereby demonstrating the safety-enhancement capabilities of the system.

B. Communication and Digitalization

Sensor and actuator data were transmitted to a Fog node (a personal computer) hosting the middleware components. This included an MQTT broker for real-time data exchange and the *Eclipse BaSyx* framework [62] for AAS-based asset representation. The *BaSyx Databridge* [63] was used to map MQTT topics to AAS properties, enabling bi-directional event-based synchronization between physical assets and their digital counterparts. The *BaSyx Databridge* also supports integration with industrial protocols such as OPC UA, Kafka [64], and PLC drivers via *Apache PLC4X* [65], ensuring interoperability with heterogeneous systems.

For digitalization, AAS types and instances were created using the *AASX Package Explorer* [66]. Figure 4 shows the Unified Modeling Language (UML) model of AAS types and instances representing the prototype. The *PhysicalDevice* type, which includes *Submodels* such as *DigitalNameplate* [67] and *TechnicalData* [50], served as the basis for creating instances of each device.

Additional AAS types were defined for *Sensor* and *Actuator*, each with *Submodels* (*MeasurementData* and *ActuationData*, respectively). *Concept Descriptions* and semantic references were included to link *Submodel* elements to global semantic identifiers, ensuring machine-readable, unambiguous interpretation [49].

Data from the AAS instances were forwarded to InfluxDB via Telegraf [68], enabling persistent time-series storage. This allowed both real-time monitoring and historical analysis of the pool environment.

C. SmartPool Web Application

A user-friendly web application (Figure 5) was developed with React [69] and Vite [70]. The interface enables real-time visualization of sensor readings, historical dashboards (via Grafana [71]), alert notifications, and direct control of actuators. Notifications from the object detection system are also integrated through MQTT.

The design emphasized both usability and system functionality, ensuring operators can access key information quickly while maintaining the ability to act on system alerts. In this way, the *SmartPool* application illustrates how AAS-based digital representations can be made accessible to human operators in an intuitive form.

Finally, Figure 6 summarizes the complete prototype architecture, showing how the physical assets, middleware, AAS environment, and application layers are integrated to form a functioning end-to-end system.

VI. Discussion of Results

The implementation and validation of the *SmartPool* system demonstrated the feasibility and practical benefits of integrating CPS and standardized digital twin frameworks in the domain of aquatic facility management. The findings from the prototype not only confirm the functional validity of the proposed architecture but also highlight key contributions to

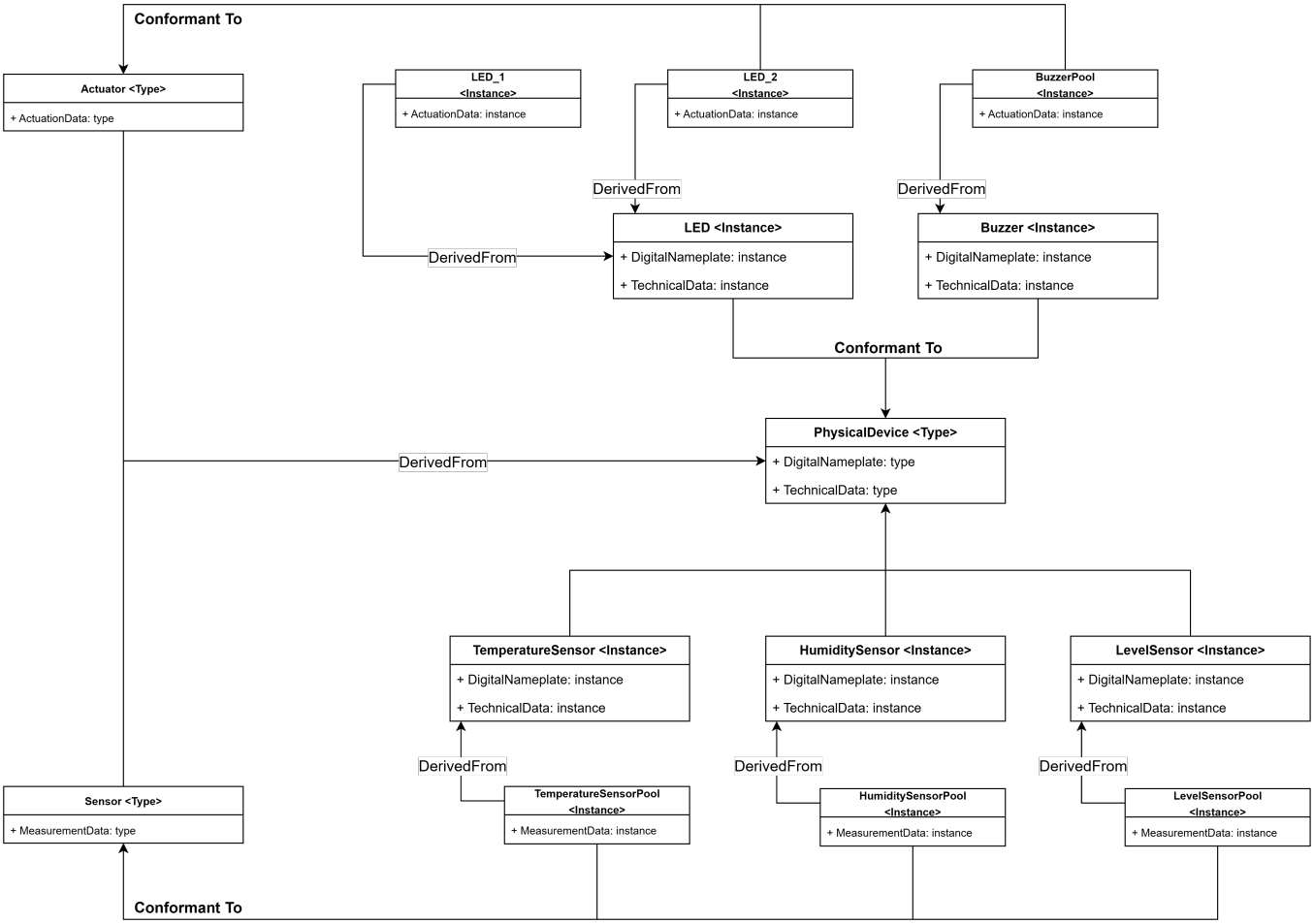


Figure 4. UML diagram of AAS types and instances for the prototype pool environment.

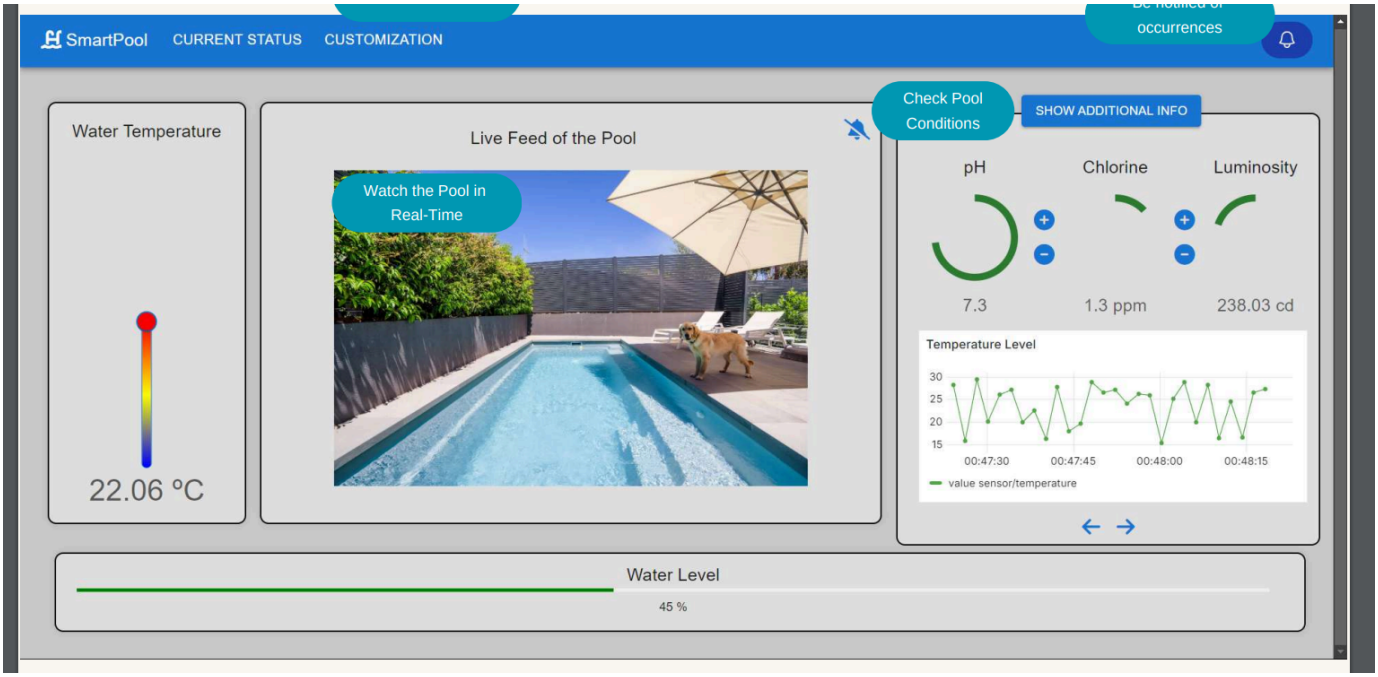


Figure 5. *SmartPool* Prototype Web Application.

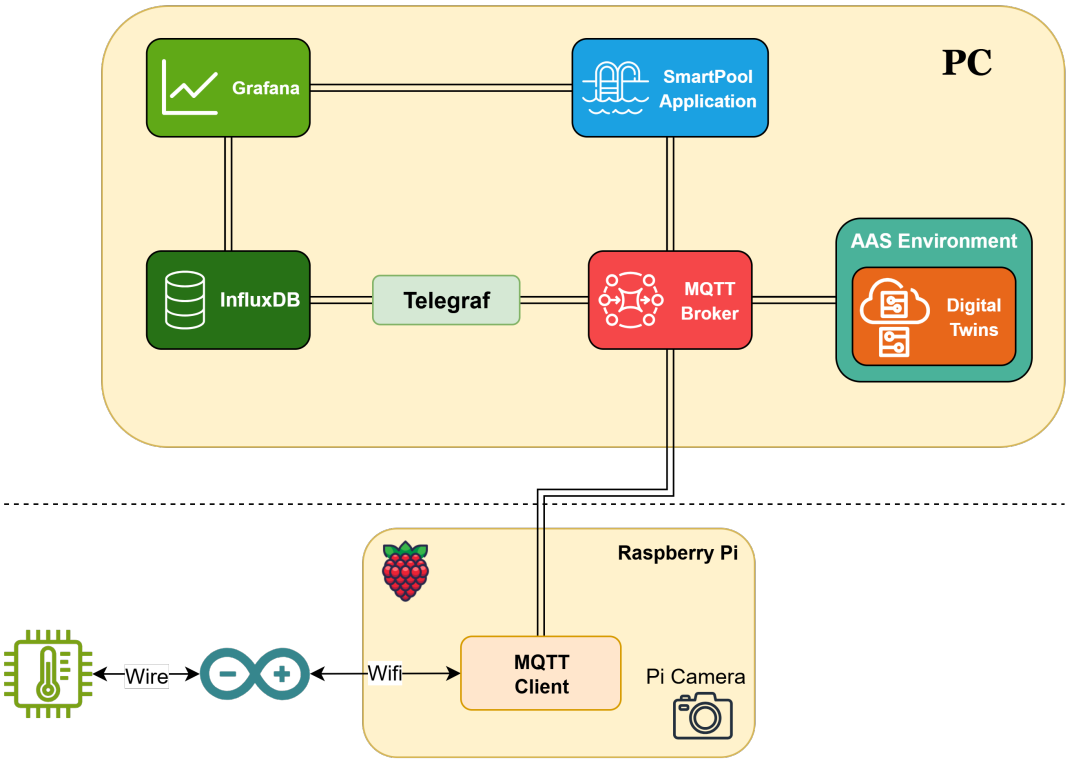


Figure 6. Prototypes System Architecture.

current knowledge and identify limitations that should be addressed in future work.

A. Insights from Prototype Validation

The experimental validation showed that the *SmartPool* architecture supports real-time sensing, actuation, and digital representation of pool components using lightweight, interoperable technologies. Key achievements include:

- Reliable data acquisition and actuation: The system demonstrated consistent and timely readings of water temperature and turbidity. Combined with automated rule-based actuation, this enables near real-time response to deviations in water quality, contributing to improved maintenance efficiency.
- Integration of visual safety monitoring: The edge-level video processing node detected human presence and movement in the pool area. While the current frame rate was limited (1 frame every 3 seconds), it proved sufficient to identify basic movement patterns for safety alerts, confirming the feasibility of edge-based drowning prevention mechanisms.
- DT representation with AAS: The deployment of AAS using the *Eclipse BaSyx* middleware enabled the standardized digital modeling of physical assets. This enhanced interoperability and reusability by allowing external applications to query asset status, historical logs, and live operational data via APIs.
- Secure communication via MQTT over TLS: The system maintained encrypted, authenticated communication between edge, fog, and cloud components, supporting secure sensor data handling in line with best practices for CPS and Industry 4.0 infrastructures.

These results support the core hypothesis that combining standardized DTs with CPS-based pool management offers a scalable, secure, and flexible solution for real-time monitoring and automation in aquatic environments.

B. Comparison with Existing Solutions

Compared to related works reviewed, *SmartPool* advances the state of the art in multiple dimensions:

- Water Quality Monitoring: Prior work (e.g., Hamid *et al.* [28] and Lakshmikantha *et al.* [29]) focused primarily on parameter tracking without integration into broader architectural frameworks. *SmartPool* extends this by embedding sensors within a layered CPS architecture, supporting actuation, lifecycle management, and cloud synchronization.
- Safety and Drowning Prevention: Unlike work such as Sangeetha *et al.* [30] and Raj *et al.* [31], which use sensor-triggered alarms or simple image capture, *SmartPool* leverages event-driven video processing at the edge, allowing more intelligent surveillance with reduced latency and bandwidth usage.
- Communication and Connectivity: Building on the insights from Christopher *et al.* [32], *SmartPool* demon-

strates effective use of MQTT over TLS within constrained aquatic environments, where traditional connectivity solutions often fail due to humidity, water reflection, or electromagnetic interference.
- Architectural Innovations: While prior IoT architectures (e.g., Glória *et al.* [33]) demonstrated the feasibility of decentralized control using Raspberry Pi and Arduino, *SmartPool* adds an additional layer of semantic interoperability via AAS, bridging the gap between raw data and contextualized asset management—a major leap toward Industry 4.0 integration.

SmartPool not only reproduces functionalities found in the literature but integrates them cohesively into a standards-based, modular, and future-proof architecture, a distinguishing feature in this domain.

C. Contributions and Limitations

This work advances both applied and theoretical understanding in the following ways:

- Demonstrates the application of AAS DTs beyond industrial production, validating its applicability in non-traditional CPS domains such as aquatic environments;
- Provides an end-to-end reference architecture for CPS-enabled aquatic management systems, combining Edge, Fog, and Cloud layers with secure data pipelines and modular control logic;
- Contributes an open and extensible prototype that can serve as a baseline for researchers exploring IoT-based safety, sustainability, and DT frameworks in smart environments;
- Offers a methodology that can be generalized to other safety-critical settings (e.g., public fountains, recreational facilities), reinforcing the role of CPS and interoperability standards in public infrastructure management.

Despite its contributions, *SmartPool* still presents several limitations:

- Limited AI Capabilities: While basic video analysis was implemented, more sophisticated ML-based behavior recognition (e.g., distress posture detection, fall prediction) requires both more powerful hardware and annotated datasets, which were outside the scope of this prototype.
- Frame Rate Constraints: The current processing pipeline limits video analysis to 3-second intervals, reducing responsiveness in fast-evolving drowning scenarios. Future versions should optimize the video subsystem or integrate hardware accelerators.
- Energy Consumption: No energy optimization was implemented at the prototype stage. Power-hungry components such as the camera and WiFi module may challenge long-term deployments in solar-powered or remote setups.
- Scalability: The system was tested with a single pool setup. Real-world scenarios with multiple pools and users may require load balancing, data deduplication, and improved concurrency handling across the middleware stack.
- Security and Identity Management: Although TLS and JSON Web Tokens (JWT) [72] were implemented, a full Public Key Infrastructure (PKI) [73] was not realized. Identity federation, certificate rotation, and fine-grained access control are planned as future enhancements.

VII. Conclusion & Future Work

This research introduced *SmartPool*, a modular and standards-based prototype platform for intelligent swimming pool management, combining CPS principles with AAS DT representations. By integrating real-time sensing, automated actuation, and edge-level visual surveillance within a secure and interoperable architecture, the system provides a novel approach to pool monitoring, safety, and lifecycle management.

The prototype validated the feasibility of applying CPS and Industry 4.0 paradigms in non-industrial contexts such as recreational aquatic facilities. Key outcomes included accurate real-time tracking of water quality parameters, responsive actuation logic, and the demonstration of digital asset models for pool components. Furthermore, the inclusion of edge-based visual monitoring reinforced the potential for enhancing safety in semi-supervised environments.

In addition to its technical achievements, *SmartPool* contributes new insights into:

- The application of AAS-based DTs beyond manufacturing environments.
- The use of lightweight, modular architectures to enable real-time control and monitoring in resource-constrained CPS deployments.
- The integration of safety-enhancing features, such as drowning detection and anomaly response, into a unified, standards-compliant middleware stack.

However, as with many academic proofs-of-concept, the system faces several known limitations. These include latency in edge-based video analysis, modest scalability, simplified environmental conditions, and the absence of fully implemented cybersecurity frameworks. These limitations underscore the challenges of balancing responsiveness, complexity, and energy efficiency in real-world deployments.

To build on these findings, several promising avenues for future research and development are proposed:

1) **Edge AI and Smart Perception:** Integrating optimized machine learning models (e.g., TinyML [74] or YOLOv5-tiny [75]) for local event recognition could improve detection capabilities while maintaining real-time responsiveness. This would enable behavior-based alarms (e.g., drowning postures, unauthorized access) and predictive maintenance based on sensor patterns.
2) **Secure and Trustworthy CPS:** Implementing a certificate-based PKI for fog and middleware nodes, along with TLS-secured MQTT communication and node-level authentication, will significantly enhance trust, integrity, and data confidentiality in distributed deployments. Additionally, availability concerns in edge scenarios must be addressed with fault-tolerant fallback routines, while data integrity mechanisms—such

as checksums, hash verification, and immutable logging—should be adopted to mitigate tampering risks.

3) **Energy Efficiency and Sustainability:** Exploring energy harvesting (e.g., solar panels) and ultra-low-power sensing technologies can reduce the environmental footprint and make the solution viable for remote or off-grid installations.
4) **Scalability and Multi-Pool Coordination:** Extending the platform to manage multiple pools simultaneously—or pools within large public/commercial facilities—will test its robustness and enable integration into broader smart building or smart city ecosystems.
5) **AAS-Integrated Analytics and Decision Support:** The AAS framework can be extended to include not only real-time parameters but also predictive indicators, alerts, and lifecycle metadata, transforming digital twins into decision-support agents rather than mere data containers.
6) **Control Loop Optimization and Adaptive Logic:** While the current system employs threshold-based actuation with fixed margins, future work could evaluate more sophisticated control strategies. These include PID or fuzzy logic control to ensure stable chemical dosing, model-predictive control (MPC) for resource optimization, and reinforcement learning approaches that adapt actuation parameters based on evolving environmental feedback.
7) **Broader Applicability in Water Environments:** The SmartPool architecture is not limited to recreational pools. Its modularity and asset-oriented design make it adaptable to other domains, such as aquaculture monitoring, thermal bath management, water purification facilities, and even environmental sensing in lakes or reservoirs. Future deployments could explore domain-specific adjustments to sensing, control, and safety logic to validate cross-context adaptability.

In conclusion, this research demonstrates the potential of integrating IoT, CPS, and AAS-based digital twins into intelligent water management systems. It contributes to both engineering practice and scientific knowledge by illustrating how lifecycle-oriented, secure, and interoperable architectures can be adopted in traditionally fragmented, non-industrial domains. As smart environments and sustainable infrastructure initiatives evolve, solutions like *SmartPool* can serve as foundational blueprints for scalable, intelligent, and context-aware water management.

Acknowledgment

This work is financially supported by national funds through the FCT/MCTES (PIDDAC), under the Associate Laboratory Advanced Production and Intelligent Systems – ARISE LA/P/0112/2020 (DOI 10.54499/LA/P/0112/2020) and the Base Funding (UIDB/00147/2020) and Programmatic Funding (UIDP/00147/2020) of the R&D Unit Center for Systems and Technologies – SYSTEC.

References

[1] A. Ávila *et al.*, "Smartpool: An automated cps-based system for real-time water quality management," in *INTELLI 2025, The Fourteenth International Conference on Intelligent Systems and Applications*, IARIA, 2025.

[2] RubyHome Blog, "Swimming pool statistics (2025)," 2025, [Online]. Available: https : / / www . rubyhome . com / blog / swimming-pool-stats/ (visited on 12/26/2025).

[3] A. Jemat, S. Yussof, S. S. Sameon, and N. A. Alya Rosnizam, "IoT-Based System for Real-Time Swimming Pool Water Quality Monitoring," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13051 LNCS, pp. 332–341, 2021, Cited by: 3. DOI: 10.1007/978-3-030-90235-3_29.

[4] A. Angdresey, L. Sitanayah, and V. J. A. Sampul, "Monitoring and Predicting Water Quality in Swimming Pools," *EPI International Journal of Engineering*, vol. 8, no. 2, pp. 119–125, 2020, Accessed from https://doi.org/10.25042/epi-ije.082020.05. DOI: 10.25042/epi-ije.082020.05.

[5] F. Tao, Q. Qi, L. Wang, and A. Nee, "Digital twins and cyber–physical systems toward smart manufacturing and industry 4.0: Correlation and comparison," *Engineering*, vol. 5, no. 4, pp. 653–661, 2019.

[6] A. Parnianifard *et al.*, "Digital-twins towards cyber-physical systems: A brief survey," *Engineering Journal*, vol. 26, no. 9, pp. 47–61, 2022.

[7] IEC-63278, *Asset Administration Shell for Industrial Applications - Part 1: Asset Administration Shell Structure*, en, International Standard, Dec. 2023. DOI: 9782832276792.

[8] M. Kaur, V. Mishra, and P. Maheshwari, "The Convergence of Digital Twin, IoT, and Machine Learning: Transforming Data into Action," in *Internet of Things*, Springer, Cham, 2020, pp. 3–17. DOI: 10.1007/978-3-030-18732-3_1.

[9] A. Redelinghuys, A. Basson, and K. Kruger, "A six-layer architecture for the digital twin: a manufacturing case study implementation," *Journal of Intelligent Manufacturing*, vol. 31, no. 6, pp. 1383–1402, 2020. DOI: 10.1007/s10845-019-01516-6.

[10] S. Zeb *et al.*, "Industrial digital twins at the nexus of NextG wireless networks and computational intelligence: A survey," *Journal of Network and Computer Applications*, vol. 200, p. 103 309, 2022, ISSN: 1084-8045. DOI: https://doi.org/10.1016/j.jnca.2021.103309.

[11] S. Karthik, D. Priya E.L., G. Anand K.R., and A. Sharmila, "IoT Based Safety Enhanced Swimming Pool with Embedded Techniques to reduce drowning accidents," in *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, 2020, pp. 843–847. DOI: 10.1109/ICOSEC49089.2020.9215247.

[12] S. Kannoth *et al.*, "Enabling smes to industry 4.0 using the basyx middleware: A case study," in *European Conference on Software Architecture*, Springer, 2021, pp. 277–294.

[13] P. Juhás and K. Molnár, "Key components of the architecture of cyber-physical manufacturing systems," *Industry 4.0*, vol. 2, no. 5, pp. 205–207, 2017.

[14] S. Ismail, D. W. Dawoud, N. Ismail, R. Marsh, and A. S. Alshami, "Iot-based water management systems: Survey and future research direction," *IEEE Access*, vol. 10, pp. 35 942–35 952, 2022.

[15] K. Joseph, A. K. Sharma, and R. Van Staden, "Development of an intelligent urban water network system," *Water*, vol. 14, no. 9, p. 1320, 2022.

[16] J. Y. Uwamungu *et al.*, "Future of water/wastewater treatment and management by industry 4.0 integrated nanocomposite

manufacturing," *Journal of Nanomaterials*, vol. 2022, no. 1, p. 5 316 228, 2022.

[17] L. García, L. Parra, J. M. Jimenez, J. Lloret, and P. Lorenz, "Iot-based smart irrigation systems: An overview on the recent trends on sensors and iot systems for irrigation in precision agriculture," *Sensors*, vol. 20, no. 4, p. 1042, 2020.

[18] M. Singh and S. Ahmed, "Iot based smart water management systems: A systematic review," *Materials Today: Proceedings*, vol. 46, pp. 5211–5218, 2021.

[19] D. G. Pivoto *et al.*, "Cyber-physical systems architectures for industrial internet of things applications in industry 4.0: A literature review," *Journal of Manufacturing Systems*, vol. 58, pp. 176–192, 2021.

[20] M. B. Yassein, M. Q. Shatnawi, S. Aljwarneh, and R. Al-Hatmi, "Internet of things: Survey and open issues of mqtt protocol," in *2017 International Conference on Engineering & MIS (ICEMIS)*, Ieee, 2017, pp. 1–6.

[21] S. Devalal and A. Karthikeyan, "Lora technology-an overview," in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, 2018, pp. 284–290.

[22] F. C. Andriulo, M. Fiore, M. Mongiello, E. Traversa, and V. Zizzo, "Edge computing and cloud computing for internet of things: A review," in *Informatics*, MDPI, vol. 11, 2024, p. 71.

[23] R. Singh, A. Gehlot, L. R. Gupta, B. Singh, and M. Swain, *Internet of things with Raspberry Pi and Arduino*. CRC Press, 2019.

[24] J. M. Marais, D. V. Bhatt, G. P. Hancke, and T. Ramotsoela, "A web-based swimming pool information and management system," in *2016 IEEE 14th International Conference on Industrial Informatics (INDIN)*, IEEE, 2016, pp. 980–985.

[25] A. Alotaibi, "Automated and intelligent system for monitoring swimming pool safety based on the iot and transfer learning," *Electronics*, vol. 9, no. 12, p. 2082, 2020.

[26] Á. de la Puente-Gil, M. de Simón-Martín, A. González-Martínez, A.-M. Diez-Suárez, and J.-J. Blanes-Peiró, "The internet of things for the intelligent management of the heating of a swimming pool by means of smart sensors," *Sensors*, vol. 23, no. 5, p. 2533, 2023.

[27] G. Simões, C. Dionísio, A. Glória, P. Sebastião, and N. Souto, "Smart system for monitoring and control of swimming pools," in *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*, IEEE, 2019, pp. 829–832.

[28] S. A. Hamid *et al.*, "IoT based Water Quality Monitoring System and Evaluation," in *2020 10th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, 2020, pp. 102–106. DOI: 10.1109/ICCSCE50387.2020.9204931.

[29] V. Lakshmikantha *et al.*, "IoT based smart water quality monitoring system," *Global Transitions Proceedings*, vol. 2, no. 2, pp. 181–186, 2021, International Conference on Computing System and its Applications (ICCSA- 2021), ISSN: 2666-285X. DOI: https://doi.org/10.1016/j.gltp.2021.08.062.

[30] A. Sangeetha *et al.*, "Smart Swimming Pool Management System (SSPMS) using IoT," in *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, 2023, pp. 840–846. DOI: 10.1109/ICIDCA56705.2023.10099729.

[31] K. J. S. Raj *et al.*, "Enhancing Pool Safety and Efficiency with an IoT Supported Monitoring System," in *2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN)*, 2023, pp. 1232–1237. DOI: 10.1109/ICPCSN58827.2023.00208.

[32] J. P. Christopher, S. D. Damayanti, and M. Suryanegara, "Investigating IoT-LoRa Technology for The Underwater System Application," in *2023 IEEE 8th International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, 2023, pp. 1–4. DOI: 10.1109/ICRAIE59459.2023.10468069.

[33] A. Glória, F. Cercas, and N. Souto, "Design and implementation of an IoT gateway to create smart environments," *Procedia Computer Science*, vol. 109, pp. 568–575, 2017, 8th International Conference on Ambient Systems, Networks and Technologies, ANT-2017 and the 7th International Conference on Sustainable Energy Information Technology, SEIT 2017, 16-19 May 2017, Madeira, Portugal, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2017.05.343.

[34] M. Pule, A. Yahya, and J. Chuma, "Wireless sensor networks: A survey on monitoring water quality," *Journal of Applied Research and Technology*, vol. 15, no. 6, pp. 562–570, 2017.

[35] W.-C. Kao, Y.-L. Fan, F.-R. Hsu, C.-Y. Shen, and L.-D. Liao, "Next-generation swimming pool drowning prevention strategy integrating ai and iot technologies," *Heliyon*, vol. 10, no. 18, 2024.

[36] N. Cameron, "Esp32 microcontroller," in *ESP32 Formats and Communication: Application of Communication Protocols with ESP32 Microcontroller*. Berkeley, CA: Apress, 2023, pp. 1–54, ISBN: 978-1-4842-9376-8. DOI: 10.1007/978-1-4842-9376-8_1.

[37] F. P. F. Domingos, A. Lotfi, I. K. Ihianle, O. Kaiwartya, and P. Machado, "Underwater communication systems and their impact on aquatic life—a survey," *Electronics*, vol. 14, no. 1, p. 7, 2024.

[38] Lalhriatpuii, Ruchi, and V. Wasson, "Comprehensive exploration of iot communication protocol: Coap, mqtt, http, lorawan and amqp," in *International Conference on Machine Learning Algorithms*, Springer, 2024, pp. 261–274.

[39] D. Rosendo, A. Costan, P. Valduriez, and G. Antoniu, "Distributed intelligence on the edge-to-cloud continuum: A systematic literature review," *Journal of Parallel and Distributed Computing*, vol. 166, pp. 71–94, 2022.

[40] M. Elgorma *et al.*, "A review of methods for detecting and preventing drowning incorporating various techniques, devices and technologies," in *2nd International Conference on Electrical Engineering and Automatic Control*, 2024, pp. 1–6.

[41] H. Cañas, J. Mula, M. Díaz-Madroñero, and F. Campuzano-Bolarín, "Implementing industry 4.0 principles," *Computers & Industrial Engineering*, vol. 158, p. 107 379, 2021.

[42] J. Vom Brocke, A. Hevner, and A. Maedche, "Introduction to design science research," in *Design science research. Cases*, Springer, 2020, pp. 1–13.

[43] K. Keshkeh, A. Jantan, K. Alieyan, and U. M. Gana, "A review on tls encryption malware detection: Tls features, machine learning usage, and future directions," in *International Conference on Advances in Cyber Security*, Springer, 2021, pp. 213–229.

[44] S. S. Salwe and K. K. Naik, "Heterogeneous Wireless Network for IoT Applications," *IETE Technical Review*, vol. 36, pp. 61–68, 2019. DOI: 10.1080/02564602.2017.1400412.

[45] A. Banks and R. Gupta, "MQTT Version 3.1.1," 2014, [Online]. Available: https://docs.oasis-open.org/mqtt/mqtt/v3.1.1/mqtt-v3.1.1.html (visited on 12/26/2025).

[46] OPC Foundation, *Opc unified architecture (opc ua)*, Accessed: 2025-12-26, 2024.

[47] IEC-62541, *OPC Unified Architecture - Part 1: Overview and concepts*, en, International Standard, Nov. 2020. DOI: 9782832290767.

[48] Apache PLC4X, *Apache PLC4X Modbus Protocol Guide*, Accessed: 2025-12-26, 2019.

[49] P. Industrie 4.0, *Details of the Asset Administration Shell - Part 1*, en, May 2022.

[50] ZVEI, *Submodel Templates of the Asset Administration Shell*, https://www.zvei.org/fileadmin/user_upload/Presse_und_Medien/Publikationen/2020/Dezember/Submodel_Templates_

of_the_Asset_Administration_Shell/201117_I40_ZVEI_SG2_Submodel_Spec_ZVEI_Technical_Data_Version_1_1.pdf, Accessed: 2025-12-26, 2020.

[51] E. Barnstedt *et al.*, "Open source drives digital twin adoption," *IIC J. Innov*, 2021.

[52] E. Tantik and R. Anderl, "Integrated data model and structure for the asset administration shell in industrie 4.0," *Procedia Cirp*, vol. 60, pp. 86–91, 2017.

[53] T. Abdel-Aty, E. Negri, and S. Galparoli, "Asset Administration Shell in Manufacturing: Applications and Relationship with Digital Twin," *IFAC-PapersOnLine*, vol. 55, pp. 2533–2538, 2022, Issue: 10. DOI: 10.1016/j.ifacol.2022.10.090.

[54] IEC-62264, *Enterprise-control system integration Part 1: Models and terminology*, en, International Standard, May 2013. DOI: 9782832208335.

[55] eCl@ss, *Ecl@ss – the standard for product and service classification*, Accessed: 2025-12-26, 2020.

[56] Eclipse Foundation, *Eclipse BaSyx Wiki*, https://wiki.basyx.org/en/latest/, Accessed: 2025-12-26, 2024.

[57] W. Quadrini, C. Cimino, T. Abdel-Aty, L. Fumagalli, and D. Rovere, "Asset Administration Shell as an interoperable enabler of Industry 4.0 software architectures: A case study," *Procedia Computer Science*, vol. 217, pp. 1794–1802, 2022. DOI: 10.1016/j.procs.2022.12.379.

[58] M. Banzi and M. Shiloh, *Getting Started with Arduino*, 3rd. Sebastopol, CA, USA: Maker Media, Inc., 2014, ISBN: 978-1449363338.

[59] M. Pagnutti *et al.*, "Laying the foundation to use raspberry pi 3 v2 camera module imagery for scientific and engineering purposes," *Journal of Electronic Imaging*, vol. 26, no. 1, pp. 013 014–013 014, 2017.

[60] OpenCV Team, "OpenCV: Open Source Computer Vision Library," 2025, [Online]. Available: https://opencv.org/ (visited on 12/26/2025).

[61] M. Sohan, T. Sai Ram, and C. V. Rami Reddy, "A review on yolov8 and its advancements," in *International Conference on Data Intelligence and Cognitive Informatics*, Springer, 2024, pp. 529–545.

[62] E. Foundation, "Eclipse BaSyx Documentation," 2023, [Online]. Available: https://www.eclipse.org/basyx/ (visited on 12/26/2025).

[63] Eclipse BaSyx, *Databridge - Eclipse BaSyx Component*, https://wiki.basyx.org/en/latest/content/user_documentation/basyx_components/databridge/index.html, Accessed: 2025-12-26, 2024.

[64] N. Garg, *Apache kafka*. Packt Publishing, 2013.

[65] Apache PLC4X, *Apache PLC4X*, https://plc4x.apache.org/, Accessed: 2025-12-26, 2017.

[66] IDTA, *AASX Package Explorer*, https://github.com/admin-shell-io/aasx-package-explorer, Accessed: 2025-12-26, 2020.

[67] ZVEI, *THE DIGITAL NAMEPLATE CONSISTENT, SUSTAINABLE, FUTURE-PROOF, NETWORKED*, https://www.zvei.org/fileadmin/user_upload/Presse_und_Medien/Publikationen/2020/November/Das_Digitale_Typenschild_-_ZVEI-Empfehlung/Digital_Nameplate.pdf, Accessed: 2025-12-26, 2020.

[68] InfluxData, "InfluxDB Documentation," 2025, [Online]. Available: https://docs.influxdata.com/influxdb/ (visited on 12/26/2025).

[69] Meta Platforms, Inc., "React: A JavaScript Library for Building User Interfaces," 2025, [Online]. Available: https://reactjs.org/ (visited on 12/26/2025).

[70] Evan You and Vite Contributors, "Vite: Next Generation Frontend Tooling," 2025, [Online]. Available: https://vitejs.dev/ (visited on 12/26/2025).

[71] S. Kirešová *et al.*, "Grafana as a visualization tool for measurements," in *2023 IEEE 5th International Conference on Modern Electrical and Energy System (MEES)*, IEEE, 2023, pp. 1–5.

[72] P. Mahindraka, "Insights of json web token," *International International Journal of Recent Technology and Engineering (IJRTE) ISSN*, pp. 2277–3878, 2020.

[73] J. Höglund, S. Lindemer, M. Furuhed, and S. Raza, "Pki4iot: Towards public key infrastructure for the internet of things," *Computers & Security*, vol. 89, p. 101 658, 2020.

[74] Y. Abadade *et al.*, "A comprehensive survey on tinyml," *IEEE Access*, vol. 11, pp. 96 892–96 922, 2023.

[75] T. Huang, M. Cheng, Y. Yang, X. Lv, and J. Xu, "Tiny object detection based on yolov5," in *Proceedings of the 2022 5th International Conference on Image and Graphics Processing*, 2022, pp. 45–50.

An Investigation of Inconsistent Expectations of Horse Racing Experts by Analyzing Horses Classified into Three Sire Line Types

Yasuhiko Watanabe
Ryukoku University
Seta, Otsu, Shiga, Japan
watanabe@rins.ryukoku.ac.jp

Hideaki Nakanishi
Ryukoku University
Seta, Otsu, Shiga, Japan
t170517@mail.ryukoku.ac.jp

Yoshihiro Okada
Ryukoku University
Seta, Otsu, Shiga, Japan
okada@rins.ryukoku.ac.jp

***Abstract*—In recent years, statistical researches often showed even experts can make mistakes although they have a wealth of knowledge and experience. In this study, we focus on horse racing experts, such as racing horse owners and trainers, and investigate whether they have inconsistent expectations on their professional issue. Using sire line, distance of races, and order of finish as clues, we analyze the 36922 horses registered with Japan Racing Association from 2010 to 2017 statistically. The results of the statistical analysis showed that horse racing experts had inconsistent expectations on the problem of which race distance they thought were favorable for horses of a certain sire line. We think this is because many horse racing experts did not consider sire lines when deciding whether to continue to enter their horses in another race of a similar distance as they do when selecting race distance.**

Keywords-decision making; expert; Thoroughbred horse; sire line; race distance.

I. Introduction

Unlike most of us, experts have a wealth of knowledge and experience. However, even experts can sometimes make mistakes. For example, in the past, baseball coaches often taught players to aim for grounders rather than fly balls. However, in recent years, statistical researches brought a new batting approach that batters should aim for big fly balls rather than grounders. The new approach, known as the "fly-ball revolution", has surprised many baseball coaches and players around the world. The reason they were surprised is because they had a firm expectation on this issue and it was incorrect. The point is that they had one expectation on one issue. A question now arises whether experts have inconsistent expectations on one issue. In this study, we focus on horse racing experts, such as racing horse owners and trainers. In order to win horse races and get the prize money, they want to find races where their horses are more likely to win.

In order to analyze horse racing experts' inconsistent expectations, we focus on sire line, distance of races, and order of finish. A sire line is a term that refers to the paternal lineage or ancestry of a horse, especially a racehorse. Many people, especially horse racing experts, often say that a sire line can indicate the potential abilities or characteristics of a horse, such as which distance races they are good at. We analyzed horses classified into one sire line type and reported that horse racing experts had inconsistent expectations on the problem of which race distance they thought were favorable for horses of the sire line type [1]. In this study, we analyze horses classified into three sire line types and discuss whether horse racing experts had inconsistent expectations on the problem of which race distance they thought were favorable for horses of these three sire line types.

The rest of this paper is organized as follows: In Section II, we survey the related works. In Section III, we survey information about racehorses and show how to collect it. In Section IV, we show how to analyze racehorse information statistically and discuss whether horse racing experts have inconsistent expectations on their professional issue. Finally, in Section V, we present our conclusions.

II. Related work

Thoroughbred horses originated from a small number of Arab, Barb, and Turk stallions and native British mares approximately 300 years ago [2]–[4]. Since then, they have been selectively bred to improve speed and stamina, and are consequently superior competitive racehorses. Wade et al. reported a high-quality draft sequence of the genome of the horse and suggested that the horse was domesticated from a relatively large number of females, but few males [5]. McGivney et al. reported that centuries of selection for favourable athletic traits among Thoroughbreds acts on genes with functions in behavior, musculoskeletal conformation, and metabolism [6]. Recently, some genomic regions were identified as a candidate region influencing racing performance in racehorses [7]. Many researchers applied statistical models to evaluate various parameters on racing performance in Thoroughbred horses [8]. Martin, Strand and Kearney reported that the most influential parameter was distance raced [9]. Cheetham et al. investigated whether both race earnings and number of race starts were associated with horse signalment (age, sex, and breed), gait, and race surface [10]. Wells, Randle and Williams investigated how temporal, behavior, and loading related factors associated with the period before the start of the race influences racehorse performance [11]. Statistical researches are conducted not only in horse racing but also in other sports, such as baseball. In recent years, statistical researches brought a new batting approach that batters should aim for big fly balls rather than grounders [12]. Kato and Yanai reported that Shohei Otani, the Japanese superstar slugger in Major League Baseball (MLB), always aims for hitting fly balls [13]. This new batting approach, the so-called "fly-ball revolution", shows that even experts may make mistakes. It

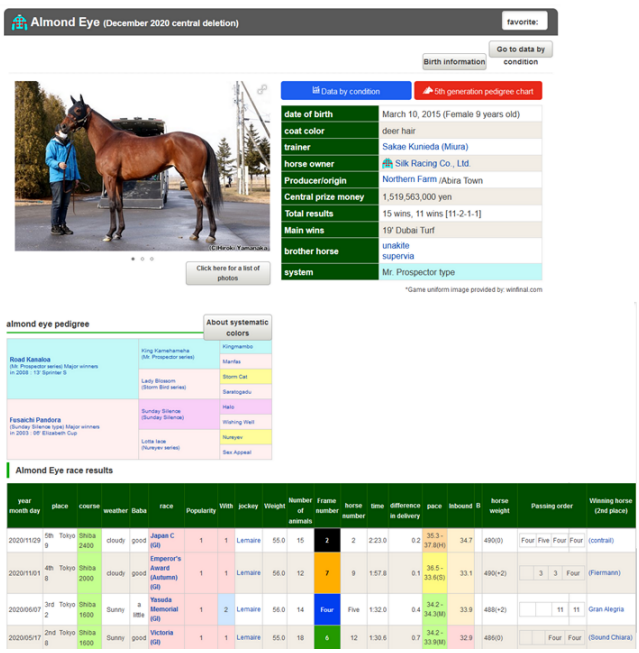


Figure 1. An example of horse information provided by Keiba Lab.

is important to discuss how and why experts made mistakes. Yerkes and Dodson studied the relationship between arousal and performance and showed that a little stress can help we perform a task, however, too much stress degrades our performance [14]. Weinschenk used Alex Rodriguez as an example to show that even experts can make mistakes when the stakes are high [15]. However, experts have a wealth of knowledge and experience, and usually have staff to share their stresses and consider issues with them. Aircraft pilots are under a great deal of mental stress when they are flying their planes. Shappell and Wiegmann focused on preventing errors in aviation, including decision errors, and propose a framework for analyzing and classifying human errors [16]. Kang and Yoon studied the types of errors that both younger and older adults make when learning how to use new technologies [17]. They found that older adults used different strategies than younger adults. However, they did not report how experts made mistakes. Bechara et al. studied unconscious mental processing and reported unconscious minds picked up danger first [18]. However, they did not study whether unconscious minds affect experts' expectations.

III. A COLLECTION OF RACEHORSE INFORMATION

Keiba Lab [19] is one of the most popular horse racing information sites in Japan. This site records various information about all racehorses registered with Japan Racing Association (JRA) and registered users can freely access it. Figure 1 shows an example of horse information provided by Keiba Lab. As shown in Figure 1, the horse information provided by Keiba Lab consists of personal information and race results. Personal data consists of name, date of birth, age, sex, coat color, breeder, birth place, owner, trainer, ancestors up to three generations ago, sire line, career statistics, career prize money, and so on. Race results consist of venue, event date, distance, weather, racetrack, surface, race name, favourite, order of finish, jockey, weight, horse number, frame number, time, and so on. In order to discuss whether horse racing experts have inconsistent expectations on their professional issue, we collected information about 36922 horses registered with JRA from 2010 to 2017 from Keiba Lab. Table I shows the number of horses registered with JRA from 2010 to 2017.

TABLE I
THE NUMBER OF HORSES REGISTERED WITH JRA FROM 2010 TO 2017.

year	number of registered horses
2010	4470
2011	4524
2012	4505
2013	4595
2014	4672
2015	4663
2016	4738
2017	4755
Total	36922

TABLE II
THE NUMBER OF HORSES CLASSIFIED INTO THE THREE MAIN SIRE LINE TYPES.

sire line	number of horses
Native Dancer Line	8799
Nearctic Line	6383
Royal Charger Line	18104
others	3636
Total	36922

On Keiba Lab, various sire lines are used to classify horses. We surveyed how racehorse sire lines diverged and grouped them into

- Native Dancer Line,
- Nearctic Line,
- Royal Charger Line, and
- others.

For example, Figure 1 shows that the sire line of *Almond Eye* was Mr. Prospector Line. It branched out from Native Dancer Line. As a result, in this study, we determined that the sire line of *Almond Eye* was Native Dancer Line. Then, we classified 36922 horses registered with JRA from 2010 to 2017 into these four types. Table II shows the number of horses classified into these four sire line types. As shown in Table II, 90 percent of the 36922 horses were classified into the three main sire lines: Native Dancer Line, Nearctic Line, and Royal Charger Line.

36922 horses had competed in races of various distances. We grouped the race distances into five types: 1000 – 1399m, 1400 – 1799m, 1800 – 2199m, 2200 – 2799m, and more than 2800m. Then, we investigated which distance races and how many times the 36922 horses had competed in. For example, *Almond Eye* had competed in one 1000–1399m race, six 1400–1799m races, four 1800–2199m races, and four 2200 – 2799m races. Table III shows the distance and number of races the

TABLE III
THE DISTANCE AND NUMBER OF RACES THE 36922 HORSES HAD COMPETED IN.

	race distance					
	1000-1399m	1400-1799m	1800-2199m	2200-2799m	2800m-	Total
number of races	5144	7433	7247	2083	1244	23151

TABLE IV
THE NUMBER OF TIMES THE 36922 HORSES OF FOUR SIRE LINES HAD COMPETED IN RACES OF VARIOUS DISTANCES.

	race distance					
sire line	1000-1399m	1400-1799m	1800-2199m	2200-2799m	2800m-	Total
Native Dancer	27008	31619	28568	4173	2511	93879
Nearctic	18710	22444	20072	2838	1647	65711
Royal Charger	42525	67514	71758	13181	5848	200826
others	9879	12058	10780	1817	876	35410
Total	98122	133635	131178	22009	10882	395826

TABLE V
THE NUMBER OF HORSES OF FOUR SIRE LINES HAD COMPETED IN RACES OF VARIOUS DISTANCES.

	race distance				
sire line	1000-1399m	1400-1799m	1800-2199m	2200-2799m	2800m-
Native Dancer	5045	7135	5599	1269	574
Nearctic	3641	5102	4053	1005	395
Royal Charger	8807	14666	13074	3794	1320
others	2056	2844	2312	561	207
Total	19549	29747	25038	6629	2496

TABLE VI
THE NUMBER OF TIMES THE 36922 HORSES OF FOUR SIRE LINES HAD FINISHED IN FIRST PLACE IN THE RACES OF VARIOUS DISTANCES.

	race distance					
sire line	1000-1399m	1400-1799m	1800-2199m	2200-2799m	2800m-	Total
Native Dancer	1947	2261	2121	341	188	6858
Nearctic	1347	1511	1399	206	143	4606
Royal Charger	2580	4767	5496	1078	495	14416
others	677	855	671	105	52	2360
Total	6551	9394	9687	1730	878	28240

TABLE VII
THE NUMBER OF TIMES THE 36922 HORSES OF FOUR SIRE LINES HAD FINISHED IN TOP FIVE PLACE IN THE RACES OF VARIOUS DISTANCES.

	race distance					
sire line	1000-1399m	1400-1799m	1800-2199m	2200-2799m	2800m-	Total
Native Dancer	9345	10912	10552	1748	1120	33677
Nearctic	6462	7700	7112	1070	728	23072
Royal Charger	13893	23937	26949	5369	2713	72861
others	3203	4054	3564	655	317	11793
Total	32903	46603	48177	8842	4878	141403

36922 horses had competed in. Table IV shows the number of times the 36922 horses of four sire lines had competed in races of various distances. Table V shows the number of horses of four sire lines had competed in races of various distances.

Horse owners get prize money when their horses place in the top five in races held by JRA. As a result, we investigated which distance races and how many times the 36922 horses of four sire lines had finished in first place and top five place in races held by JRA. Tables VI and VII show the number of times the 36922 horses of four sire lines had finished in first place and top five place in the races of various distances, respectively.

IV. ANALYSIS OF INCONSISTENT EXPECTATIONS OF HORSE RACING EXPERTS

Horse racing experts have the problem of which distance races are favorable or unfavorable for racehorses of a certain sire line. Also, they have expectations on this problem. In this section, we investigate whether horse racing experts have inconsistent expectations on this problem.

A. Basic idea

It is widely recognized that inherited variation in physical and physiological characteristics is responsible for variation in individual aptitude for race distance. Many horse racing experts would agree that if the best race distance of ancestors is known, the offspring's best race distance is most likely to take after them. As a result, we focus on three factors of racehorses:

- sire line,
- race distance, and
- order of finish.

In this section, we first investigate whether horse racing experts entered their horses of certain sire lines in races of certain distances too many times or too few times. The result of this investigation shows which distance races the experts thought were favorable or unfavorable for racehorses of a certain sire line. Then, we investigate whether horses of certain sire lines won or lost races of certain distances too many times. The result of this investigation shows which distance races were fovorable or unfavorable for racehorses of a certain sire line. Next, we investigate whether horse racing experts entered their horses into races of a certain distance too many times. The result of this investigation shows experts' judgements of horses' performance. Finally, we compare the results of statistical analyses on experts' race selection, the race results, and experts' judgements of horses' performance, and detect inconsistent expectations of horse racing experts.

B. Detection of race distance and sire line combinations that horse racing experts selected too many times or too few times

In order to detect cases where horse racing experts entered their horses of certain sire lines into races of certain distances too many times or too few times, we conduct the statistical analysis by using Hypothesis *ES*.

Hypothesis *ES* If experts did not enter too many times or too few times their racehorses of certain sire lines into races of certain distances, we would expect that experts entered their horses of sire line s_i into races of distance d_j at most

TABLE VIII
THE P-VALUES OF EXPERTS' RACE SELECTIONS FOR HORSES OF NATIVE DANCER LINE, NEARCTIC LINE, AND ROYAL CHARGER LINE.

sire line	race distance				
	1000–1399m	1400–1799m	1800–2199m	2200–2799m	2800m–
Native Dancer	1.0000	0.3024	0.0000	0.0000	0.0825
Nearctic	1.0000	0.9835	0.0000	0.0000	0.0001
Royal Charger	0.0000	0.0882	1.0000	1.0000	0.9999

$N_{ES}(s_i, d_j)$ times

$$N_{ES}(s_i, d_j) = P_{ES}(d_j) \times \sum_j N_{entry}(s_i, d_j) \quad (1)$$

where d_j is the type of race distance. We classified race distances into five types:

d_1 1000 – 1399m
d_2 1400 – 1799m
d_3 1800 – 2199m
d_4 2200 – 2799m
d_5 2800m –

$N_{entry}(s_i, d_j)$ is the number of times horses of sire line s_i were entered into races of distance d_j, as a result, $\sum_j N_{entry}(s_i, d_j)$ is the total number of times horses of sire line s_i were entered into races. $P_{ES}(d_j)$ is the probability that an expert enters his/her horse into a race of distance d_j. $P_{ES}(d_j)$ is

$$P_{ES}(d_j) = \frac{\sum_i N_{entry}(s_i, d_j)}{\sum_i \sum_j N_{entry}(s_i, d_j)} \quad (2)$$

where $\sum_i N_{entry}(s_i, d_j)$ is the total number of times horses were entered into races of distance d_j and $\sum_i \sum_j N_{entry}(s_i, d_j)$ is the total number of times horses were entered into races.

If this hypothesis is rejected by an two-sided binomial test [20], we determine that experts entered their horses of sire lines s_i into races of distance d_j too many times or too few times.

C. Detection of race distance and sire line combinations that gave good or poor results for racehorse experts too many times

In order to detect cases where horses of certain sire lines won or lost races of certain distances too many times, we conduct the statistical analysis by using Hypothesis *RR*.

Hypothesis *RR* If horses of certain sire lines did not perform well too many times or too few times in races of certain distances, we would expect that horses of sire line s_i finished within $rank$-th place in races of distance d_j at most $N_{RR}(s_i, d_j, rank)$ times

$$N_{RR}(s_i, d_j, rank) = P_{RR}(d_j, rank) \times N_{entry}(s_i, d_j) \quad (3)$$

TABLE IX
THE P-VALUES OF RACE RESULTS OF HORSES OF NATIVE DANCER LINE, NEARCTIC LINE, AND ROYAL CHARGER LINE.

(a) Native Dancer Line

result	race distance				
	1000–1399m	1400–1799m	1800–2199m	2200–2799m	2800m–
first place	0.9997	0.8036	0.6069	0.7820	0.1506
top five place	0.9999	0.0890	0.7712	0.9542	0.1472

(b) Nearctic Line

result	race distance				
	1000–1399m	1400–1799m	1800–2199m	2200–2799m	2800m–
first place	0.9978	0.0412	0.0123	0.1230	0.8318
top five place	0.9982	0.0381	0.0001	0.0013	0.1528

(c) Royal Charger Line

result	race distance				
	1000–1399m	1400–1799m	1800–2199m	2200–2799m	2800m–
first place	0.0000	0.6279	0.9975	0.9145	0.8717
top five place	0.0001	0.9992	1.0000	0.9298	0.9889

where d_j is the type of race distance. We classified race distances into five types in the same way that we did in Hypothesis *ES*. $N_{entry}(s_i, d_j)$ is the number of times horses of sire line s_i were entered into races of distance d_j. $P_{RR}(d_j, rank)$ is the probability that a horse finished within $rank$-th place in a race of distance d_j. $P_{RR}(d_j, rank)$ is

$$P_{RR}(d_j, rank) = \frac{\sum_i N_{result}(s_i, d_j, rank)}{\sum_i N_{entry}(s_i, d_j)} \quad (4)$$

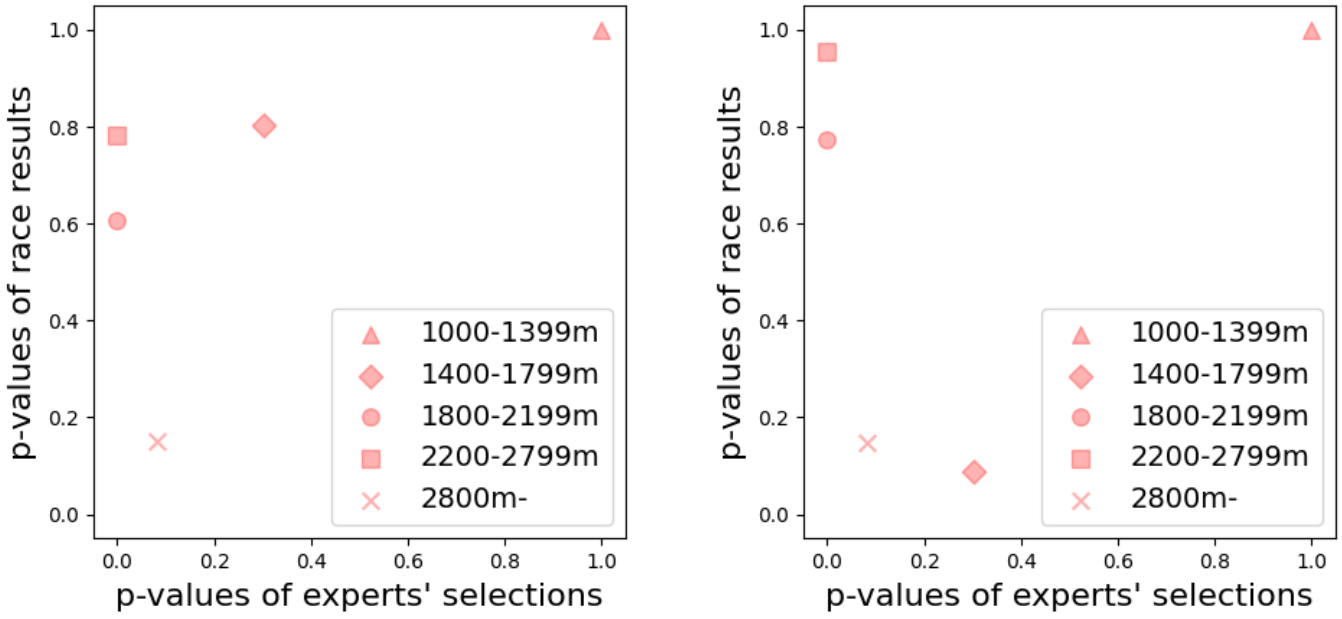
where $N_{result}(s_i, d_j, rank)$ is the number of times horses of sire line s_i finished within $rank$-th place in races of distance d_j. As a result, $\sum_i N_{result}(s_i, d_j, rank)$ is the total number of times horses finished within $rank$-th place in races of distance d_j. Furthermore, $\sum_i N_{entry}(s_i, d_j)$ is the total number of times horses were entered into races of distance d_j.

If this hypothesis is rejected by an two-sided binomial test, we determine that horses of sire line s_i finished too many times or too few times within $rank$-th place in races of distance d_j.

D. Detection of horses that horse racing experts judged to have performed well

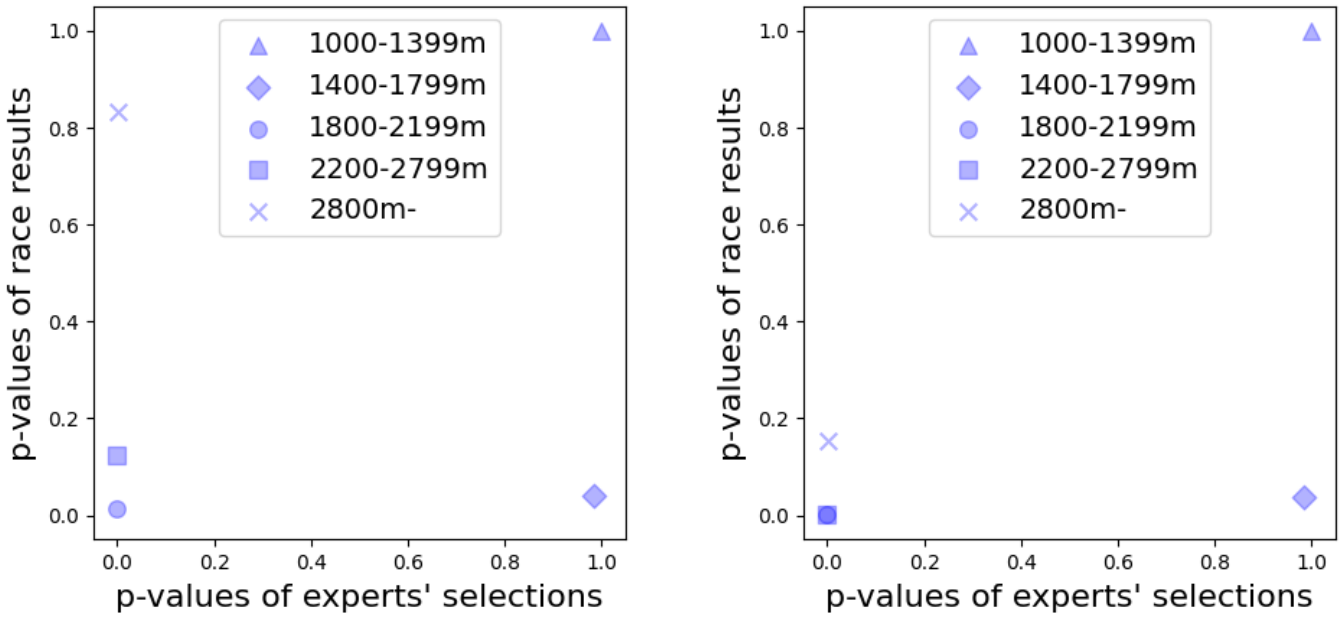
If a horse perform well in a race of a certain distance, experts will try to enter the horse into another race of a similar distance. As a result, if horses are judged to have performed well in races of a certain distance, experts may enter them into races of a similar distance repeatedly. In order to detect cases where horse racing experts entered their horses into races of certain distances too many times or too few times, we conduct the statistical analysis by using Hypothesis *EJ*.

Hypothesis *EJ* If an expert did not enter too many times or too few times his/her racehorse of a certain sire line into races



(a) first place (b) top five place

Figure 2. The p-values of experts' race selections vs race results (Native Dancer Line).



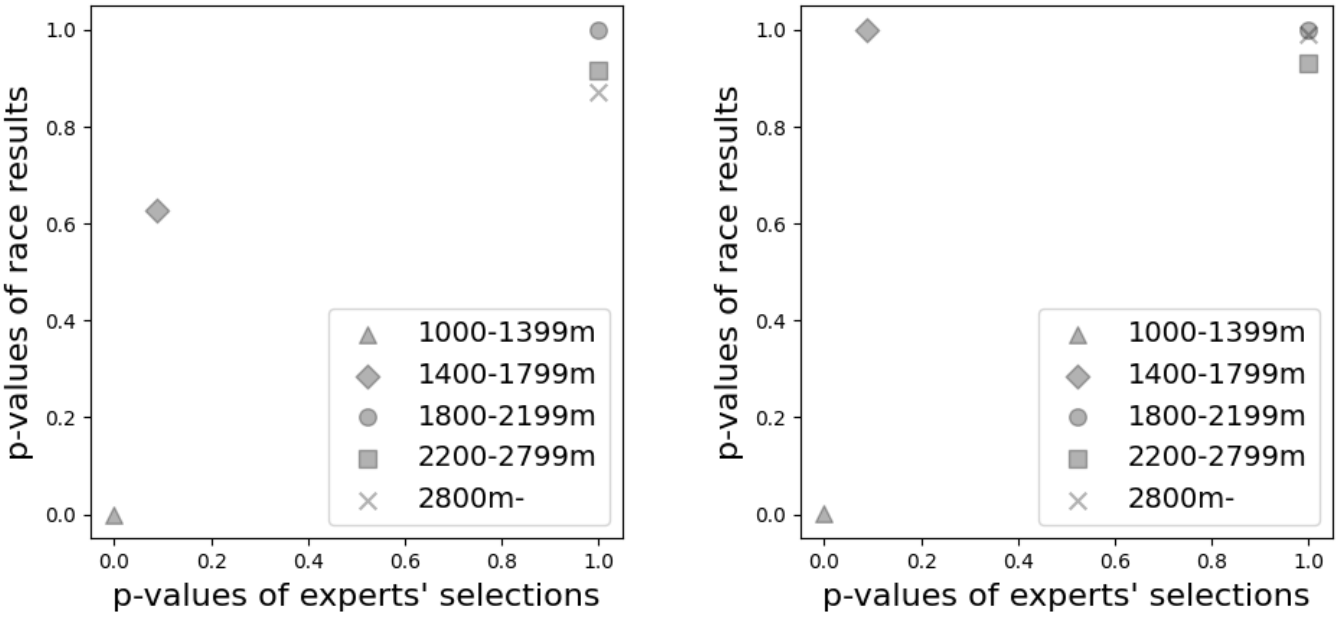
(a) first place (b) top five place

Figure 3. The p-values of experts' race selections vs race results (Nearctic Line).

of a certain distance, we would expect that the expert entered horse h_k into races of distance d_j at most $M_{EJ}(h_k, d_j)$ times

$$M_{EJ}(h_k, d_j) = P_{EJ}(s_i, d_j) \times \sum_j M_{entry}(h_k, d_j) \quad (5)$$

where s_i is the sire line of horse h_k and d_j is the type of race distance. We classified race distances into five types in the same way that we did in Hypothesis *ES*. $\sum_j M_{entry}(h_k, d_j)$ is the number of times horse h_k were entered into races. $P_{EJ}(s_i, d_j)$ is the probability that an expert enters a horse



(a) first place

(b) top five place

Figure 4. The p-values of experts' race selections vs race results (Royal Charger Line).

of sire line s_i into a race of distance d_j. $P_{EJ}(s_i, d_j)$ is

$$P_{EJ}(s_i, d_j) = \frac{N_{entry}(s_i, d_j)}{\sum_j N_{entry}(s_i, d_j)} \quad (6)$$

where $N_{entry}(s_i, d_j)$ is the number of times horses of sire line s_i were entered into races of distance d_j. As a result, $\sum_i N_{entry}(s_i, d_j)$ is the total number of times horses were entered into races of distance d_j.

If this hypothesis is rejected by an two-sided binomial test [20], we determine that an expert entered his/her horse h_k of sire lines s_i into races of distance d_j too many times or too few times.

E. Results of the investigation

In order to investigate whether horse racing experts have inconsistent expectations, we apply Hypothesis *ES*, *RR*, and *EJ* tests on

- the 8799 horses of Native Dancer Line,
- the 6383 horses of Nearctic Line, and
- the 18104 horses of Royal Charger Line

registered with JRA from 2010 to 2017, as shown in Table I. The significance levels for Hypothesis *ES*, *RR*, and *EJ* were 0.05. First, we calculated the p-values of experts' race selections, the race results, and experts' judgements of horses' performance by applying Hypothesis *ES*, *RR*, and *EJ*, respectively. Table VIII shows the p-values of experts' race selections for horses of Native Dancer Line, Nearctic Line, and Royal Charger Line. Table IX shows the p-values of race results (first place and top five place) of horses of Native Dancer Line, Nearctic Line, and Royal Charger Line. Figure 2, Figure 3, and Figure 4 show the p-values of experts' race selections vs the race results (first place and top five place) for horses of Native Dancer Line, Nearctic Line, and Royal Charger Line, respectively. Table X and Table XI show the number of

- horses determined by Hypothesis *EJ* to have repeatedly competed in races of various distances, and
- horses determined by Hypothesis *EJ* not to have repeatedly competed in races of various distances,

respectively, and also show

- the number of times the horses had competed in the races, and
- the average number of races the horses had competed in.

Figure 5 shows the percentage of horses competed repeatedly in races of various distances.

First, we consider experts' expectations that their race selections suggested. Table VIII, the results obtained by applying Hypothesis *ES*, shows

- in the case of Native Dancer Line, the p-value of race distance type d_1 (1000 – 1399m) was more than 0.975. As a result, experts entered horses of Native Dancer Line into races of distance type d_1 (1000 – 1399m) too many times. In other words, many experts strongly thought horses of Native Dancer Line were favorable to win in races of distance type d_1 (1000 – 1399m). On the other hand, the p-values of race distance type d_3 (1800 – 2199m) and d_4 (2200 – 2399m) were less than 0.025. In addition, the p-value of race distance type d_5 (2800m –) was low, 0.0825. As a result, many experts strongly thought horses

TABLE X
THE NUMBER OF HORSES DETERMINED BY HYPOTHESIS *EJ* TO HAVE REPEATEDLY COMPETED IN RACES OF VARIOUS DISTANCES AND THE AVERAGE NUMBER OF TIMES THE HORSES HAD COMPETED IN THE RACES.

(a) Native Dancer Line

	race distance				
	1000-1399m	1400-1799m	1800-2199m	2200-2799m	2800m-
horses competed repeatedly	2320	1575	1940	628	376
races competed in	20005	13491	18207	3308	2248
ave. of races competed in	8.6	8.6	9.4	5.3	6.0

(b) Nearctic Line

	race distance				
	1000-1399m	1400-1799m	1800-2199m	2200-2799m	2800m-
horses competed repeatedly	1616	1161	1427	562	274
races competed in	13709	9651	12700	2258	1506
ave. of races competed in	8.5	8.3	8.9	4.0	5.5

(c) Royal Charger Line

	race distance				
	1000-1399m	1400-1799m	1800-2199m	2200-2799m	2800m-
horses competed repeatedly	3973	3403	4445	1637	879
races competed in	32362	30280	42264	9933	5268
ave. of races competed in	8.1	8.9	9.5	6.1	6.0

TABLE XI
THE NUMBER OF HORSES DETERMINED BY HYPOTHESIS *EJ* NOT TO HAVE REPEATEDLY COMPETED IN RACES OF VARIOUS DISTANCES AND THE AVERAGE NUMBER OF TIMES THE HORSES HAD COMPETED IN THE RACES.

(a) Native Dancer Line

	race distance				
	1000-1399m	1400-1799m	1800-2199m	2200-2799m	2800m-
horses competed not repeatedly	2725	5560	3659	641	198
races competed in	7003	18128	10361	865	263
ave. of races competed in	2.6	3.3	2.8	1.3	1.3

(b) Nearctic Line

	race distance				
	1000-1399m	1400-1799m	1800-2199m	2200-2799m	2800m-
horses competed not repeatedly	2025	3941	2626	443	121
races competed in	5001	12793	7372	580	141
ave. of races competed in	2.5	3.2	2.8	1.3	1.2

(c) Royal Charger Line

	race distance				
	1000-1399m	1400-1799m	1800-2199m	2200-2799m	2800m-
horses competed not repeatedly	4834	11263	8629	2157	441
races competed in	10163	37234	29494	3248	580
ave. of races competed in	2.1	3.3	3.4	1.5	1.3

of Native Dancer Line were unfavorable to win in races over 1800m.

- in the case of Nearctic Line, the p-value of race distance type d_1 (1000 – 1399m) and d_2 (1400 – 1799m) were more than 0.975. As a result, many experts strongly thought horses of Nearctic Line were favorable to win in under 1800m races. On the other hand, the p-values of race distance type d_3 (1800 – 2199m), d_4 (2200 – 2399m), and d_5 (2800m –) were less than 0.025. As a result, many experts strongly thought horses of Nearctic Line were unfavorable to win in races over 1800m.
- in the case of Royal Charger Line, the p-value of race distance type d_3 (1800 – 2199m), d_4 (2200 – 2399m), and d_5 (2800m –) were more than 0.975. As a result, many experts strongly thought horses of Royal Charger Line were favorable to win in races over 1800m. On the other hand, the p-values of race distance type d_1 (1000 – 1399m) were less than 0.025. In addition, the p-value of race distance type d_2 (1400 – 1799m) was low, 0.0890. As a result, many experts strongly thought horses of Royal Charger Line were unfavorable to win in races under 1800m.

Next, we consider experts' expectations that the average number of races competed by horses suggested. Figure 6 shows the average number of races competed by

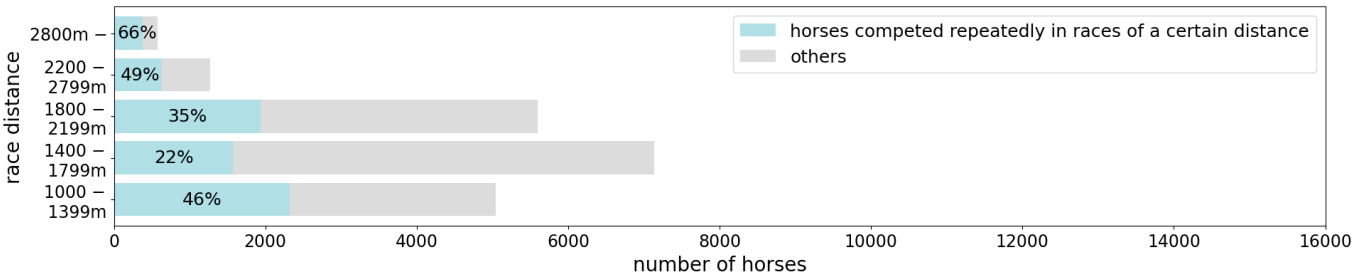
- horses determined by Hypothesis *EJ* to have repeatedly competed in races of a certain distance and
- horses determined by Hypothesis *EJ* not to have repeatedly competed in races of that distance.

For example, in the case of Native Dancer Line, the average number of races of distance type d_5 (2800m –) competed by horses determined not to have repeatedly competed in races of that distance was 1.3. Figure 6 shows that, in all three sire lines, the average number of races of distance type d_4 (2200 – 2399m) and d_5 (2800m –), in other words, the average number of races over 2200m competed by horses determined not to have repeatedly competed in races of that distance was 1.5 or less. On the other hand, the average number of races under 2200m competed by horses determined not to have repeatedly competed in races of that distance was 2.1 or more. As a result, experts had to decide whether or not to continue to enter their horses in races over 2200m based on the results of almost one race, regardless of sire line. On the other hand, they could decide whether or not to continue to enter their horses in races under 2200m based on the results of two races or more. We thought that the reason for this difference was that the number of races over 2200m was small compared to that of races under 2200m. Table III shows that races over 2200m accounted for 14% of all races. As a result, we focus on races under 2200m.

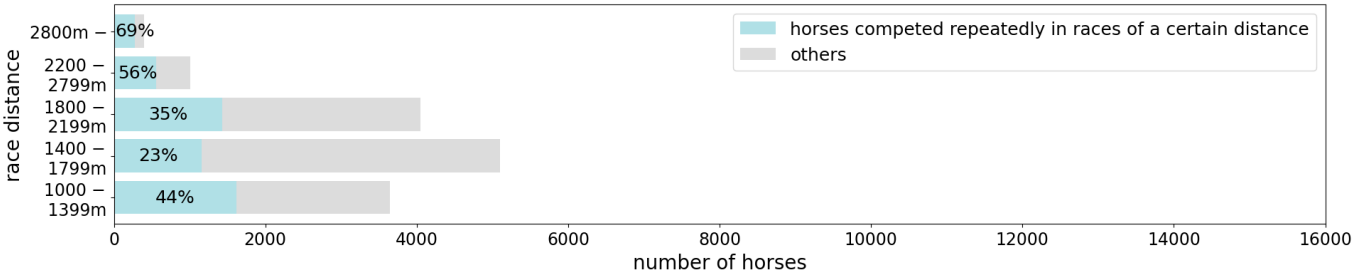
First, we consider horses of Royal Charger Line determined not to have repeatedly competed in races under 2200m. Figure 6 shows that

- the average number of races of distance type d_1 (1000 – 1399m) was 2.1.
- the average number of races of distance type d_3 (1800 – 2199m) was 3.4.

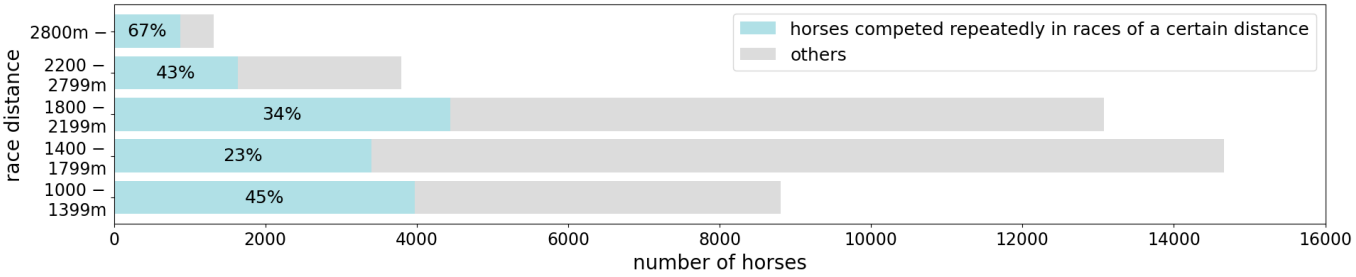
The reason for this difference is thought to be that many experts want to decide whether or not to continue to enter their horses in unfavorable races based on the results of as



(a) Native Dancer Line



(b) Nearctic Line



(c) Royal Charger Line

Figure 5. The percentage of horses competed repeatedly in races of various distances.

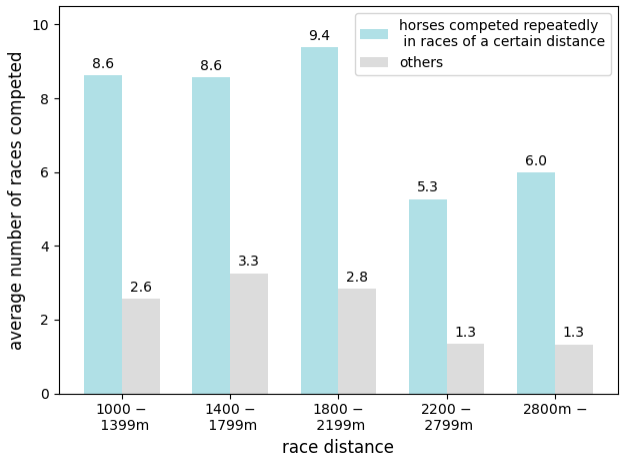
few races as possible. As mentioned, many experts thought that horses of Royal Charger Line were unfavorable to win in races of distance type d_1 (1000 – 1399m), on the other hand, they were favorable to win in races of distance type d_3 (1800 – 2199m). As a result, we thought that the average number of races at which experts decide whether to continue to enter their horses in races of distance type d_1 (1000 – 1399m) was fewer than that in races of distance type d_3 (1800 – 2199m). The point to note is the average number of races of distance type d_2 (1400 – 1799m). As shown in Figure 6, it was 3.3 and almost the same as the average number of races of distance type d_3 (1800 – 2199m). However, many experts thought that horses of Royal Charger Line were unfavorable to win in races of distance type d_2 (1400 – 1799m), on the other hand, they were favorable to win in races of distance type d_3 (1800 – 2199m). As a result, we thought that, in the case of Royal Charger Line, experts' expectations were inconsistent for races of distance type d_2 (1400 – 1799m).

Next, we consider horses of Native Dancer Line and Nearctic Line determined not to have repeatedly competed in races under 2200m. Figure 6 shows that, in both cases, the average number of races of distance type d_1 (1000 – 1399m) was almost the same as that of races of distance type d_3 (1800 – 2199m). However, horses of Native Dancer Line and Nearctic Line were favorable to win in races of distance type d_1 (1000 – 1399m), on the other hand, they were unfavorable to win in races of distance type d_3 (1800 – 2199m). As a result, we thought that, in the case of Native Dancer Line and Nearctic Line, experts' expectations were inconsistent for races under 2200m.

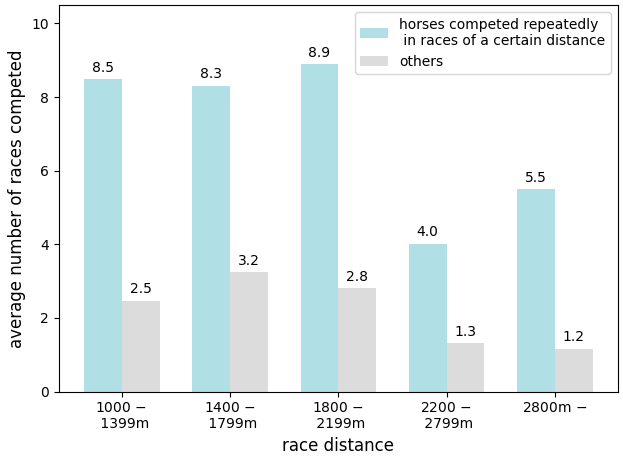
We thought the reason for this inconsistent expectations is that many experts did not consider sire lines when deciding whether to continue to enter their horses in another race of a similar distance as they do when selecting race distance. It suggests that statistical analysis may be able to resolve experts' inconsistent expectations and improve their performance.

V. Conclusion

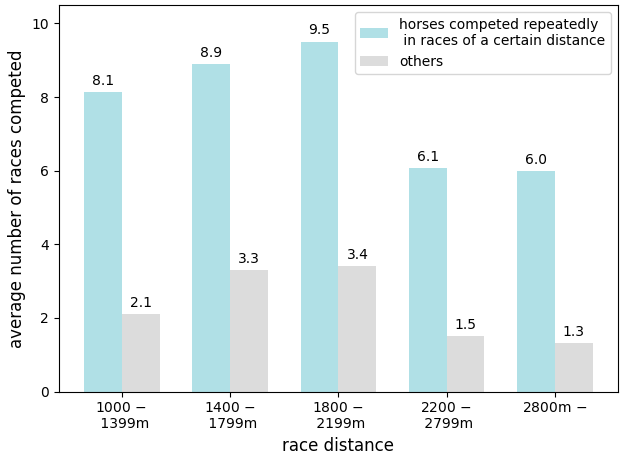
Although experts have a wealth of knowledge and experience, they sometimes make mistakes. However, not enough



(a) Native Dancer Line



(b) Nearctic Line



(c) Royal Charger Line

Figure 6. The average number of times horses determined by Hypothesis *EJ* to have and not have repeatedly competed in races of various distances.

research has been done on how and why experts made mistakes. We thought that one of the reasons why they made mistakes is that they have inconsistent expectations. As a result, in this paper, we investigated whether horse racing experts have inconsistent expectations on their professional issue. We analyzed sire lines, race distances, and race results of the 36922 horses statistically and showed that horse racing experts had inconsistent expectations on the problem of which race distance they thought were favorable for horses of a certain sire line. We think this is because many horse racing experts did not consider sire lines when deciding whether to continue to enter their horses in another race of a similar distance as they do when selecting race distance. As a result, statistical analysis may be able to resolve experts' inconsistent expectations and improve their performance.

To generalize this finding, we intend to analyze race performance data from other time periods and compare the results with those obtained in this study.

REFERENCES

[1] Y. Watanabe, H. Nakanishi, and Y. Okada, "An Investigation of Inconsistent Expectations of Horse Racing Experts," in *The Eleventh International Conference on Human and Social Analytics (HUSO 2025)*, Mar 2025, pp. 12–17. [Online]. Available: https://www.thinkmind.org/index.php?view=article&articleid=huso_2025_1_30_80018 [accessed: 2025-12-01]

[2] M. A. Bower *et al.*, "The cosmopolitan maternal heritage of the Thoroughbred racehorse breed shows a significant contribution from British and Irish native mares," *Biology Letters*, vol. 7, no. 2, pp. 316–320, 2011. [Online]. Available: https://doi.org/10.1098/rsbl.2010.0800 [accessed: 2025-12-01]

[3] E. Cunningham, J. J. Dooley, R. Splan, and D. Bradley, "Microsatellite diversity, pedigree relatedness and the contributions of founder lineages to thoroughbred horses," *Animal Genetics*, vol. 32, no. 6, pp. 360–364, 2001. [Online]. Available: https://doi.org/10.1046/j.1365-2052.2001.00785.x [accessed: 2025-12-01]

[4] E. Hill *et al.*, "History and integrity of thoroughbred dam lines revealed in equine mtDNA variation," *Animal Genetics*, vol. 33, no. 4, pp. 287–294, 2002. [Online]. Available: https://doi.org/10.1046/j.1365-2052.2002.00870.x [accessed: 2025-12-01]

[5] C. Wade *et al.*, "Genome Sequence, Comparative Analysis, and Population Genetics of the Domestic Horse," *Science*, vol. 326, no. 5954, pp. 865–867, 2009. [Online]. Available: https://doi.org/10.1126/science.1178158 [accessed: 2025-12-01]

[6] B. A. McGivney *et al.*, "Genomic inbreeding trends, influential sire lines and selection in the global Thoroughbred horse population," *Scientific Reports*, vol. 10, no. 1, p. 466, 2020. [Online]. Available: https://doi.org/10.1038/s41598-019-57389-5 [accessed: 2025-12-01]

[7] E. W. Hill, B. A. McGivney, J. Gu, R. Whiston, and D. E. MacHugh, "A genome-wide SNP-association study confirms a sequence variant (g. 66493737C>T) in the equine myostatin (MSTN) gene as the most powerful predictor of optimum racing distance for Thoroughbred racehorses," *BMC Genomics*, vol. 11, no. 1, pp. 1–10, 2010. [Online]. Available: https://doi.org/10.1186/1471-2164-11-552 [accessed: 2025-12-01]

[8] C. Wylie and J. Newton, "A systematic literature search to identify performance measure outcomes used in clinical studies of racehorses," *Equine Veterinary Journal*, vol. 50, no. 3, pp. 304–311, 2018. [Online]. Available: https://doi.org/10.1111/evj.12757 [accessed: 2025-12-01]

[9] G. Martin, E. Strand, and M. Kearney, "Use of statistical models to evaluate racing performance in thoroughbreds," *Journal of the American Veterinary Medical Association*, vol. 209, no. 11, pp. 1900–1906, 1996. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/8944806/ [accessed: 2025-12-01]

[10] J. Cheetham, A. Riordan, H. Mohammed, C. McIlwraith, and L. Fortier, "Relationships between race earnings and horse age, sex, gait, track surface and number of race starts for Thoroughbred and Standardbred racehorses in North America," *Equine Veterinary Journal*, vol. 42, no. 4,

pp. 346–350, 2010. [Online]. Available: https://doi.org/10.1111/j.2042-3306.2010.00032.x [accessed: 2025-12-01]

[11] I. Wells, H. Randle, and J. M. Williams, "Does the start of flat races influence racehorse race performance?" *Applied Animal Behaviour Science*, vol. 253, p. 105682, 2022. [Online]. Available: https://doi.org/10.1016/j.applanim.2022.105682 [accessed: 2025-12-01]

[12] T. Sawchik, *Has the Fly-Ball Revolution Begun?*, FanGraphs Baseball, 2017. [Online]. Available: https://blogs.fangraphs.com/has-the-fly-ball-revolution-begun/ [accessed: 2025-12-01]

[13] M. Kato and T. Yanai, "Launch fly balls for better batting statistics: Applicability of "fly-ball revolution" to Japan's professional baseball league," *International Journal of Performance Analysis in Sport*, vol. 22, no. 3, pp. 437–453, 2022. [Online]. Available: https://doi.org/10.1080/24748668.2022.2075302 [accessed: 2025-12-01]

[14] R. M. Yerkes and J. D. Dodson, "The relation of strength of stimulus to rapidity of habit-formation," *Journal of Comparative Neurology and Psychology*, vol. 18, no. 5, pp. 459–482, 1908. [Online]. Available: https://doi.org/10.1002/cne.920180503 [accessed: 2025-12-01]

[15] S. Weinschenk, *100 Things Every Designer Needs to Know About People, 2nd Edition*. New Riders Publishing, Jun. 2020.

[16] S. A. Shappell and D. A. Wiegmann, *The Human Factors Analysis and Classification System–HFACS*, U.S. Department of Transportation Federal Aviation Administration, 2000. [Online]. Available: https://www.skybrary.aero/sites/default/files/bookshelf/1481.pdf [accessed: 2025-12-01]

[17] N. E. Kang and W. C. Yoon, "Age- and experience-related user behavior differences in the use of complicated electronic devices," *International Journal of Human-Computer Studies*, vol. 66, no. 6, pp. 425–437, 2008. [Online]. Available: https://doi.org/10.1016/j.ijhcs.2007.12.003 [accessed: 2025-12-01]

[18] A. Bechara, H. Damasio, D. Tranel, and A. R. Damasio, "Deciding Advantageously Before Knowing the Advantageous Strategy," *Science*, vol. 275, no. 5304, pp. 1293–1295, 1997. [Online]. Available: https://doi.org/10.1126/science.275.5304.1293 [accessed: 2025-12-01]

[19] *Keiba Lab*, Keiba Lab. [Online]. Available: https://www.keibalab.jp/ [accessed: 2025-12-01]

[20] M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods, 2nd Edition*. Wiley-Interscience, Jan. 1999.

An In-Depth Analysis of a Multi-Sensor System for Smart City Road Maintenance: Detailed Design, Implementation, and Validation of a LiDAR and AI-Driven Approach

Giovanni Nardini*, Roberto Nucera*, Alessandro Ulleri*, Stefano Cordiner†,
Eugenio Martinelli†, Arianna Mencattini† and Iulian Gabriel Coltea*
*Key To Business s.r.l., Rome, Italy
e-mail: g.nardini@key2.it, r.nucera@key2.it, a.ulleri@key2.it, i.coltea@key2.it
†Department of Industrial Engineering, University of Rome Tor Vergata, Rome, Italy
e-mail: stefano.cordiner@uniroma2.eu, eugenio.martinelli@uniroma2.eu, mencattini@eln.uniroma2.it

***Abstract*—This paper details a complete, vehicle-mounted system for automated road surface inspection, developed to enhance the efficiency and safety of large-scale urban infrastructure management. As an extended version of our previous work presented at the SMART 2025 conference, this study provides an in-depth analysis of the system architecture, a refined Artificial Intelligence (AI) pipeline, and a detailed performance evaluation under challenging real-world scenarios. The system operates on a multi-sensor fusion principle, integrating High-Resolution Light Detection and Range (LiDAR) point clouds for precise 3D geometry, camera imagery for visual texture analysis, and high-accuracy Global Navigation Satellite Systems (GNSS) and inertial data for robust georeferencing. Its AI capabilities are driven by custom models: a fine-tuned Convolutional Neural Networks (CNNs) model detects and classifies road defects like potholes and cracks in images, while a Visual Transformer (ViT) semantic segmentation model provides comprehensive semantic scene understanding to avoid false positives. Through a precise LiDAR-camera calibration, these 2D detections are then projected into the 3D domain of the point clouds. This critical step isolates each defect, allowing for the creation of a three-dimensional model and the precise quantification of its physical properties, such as surface area, depth, and volume. A significant contribution of this work is the extensive validation conducted across dozens of kilometers in a complex urban road environment in Rome, Italy. We present key quantitative results that achieve high detection accuracy and centimeter-level measurement precision. Furthermore, we discuss the iterative tuning process that overcame operational challenges like motion blur, misclassification of manholes, shadows, and road markings. The findings confirm that the system is a robust and scalable solution, with a pipeline optimized for edge computing to enable real-time analysis, delivering actionable data through a map-based web portal to facilitate proactive urban road management.**

Keywords-smart cities; road maintenance; LiDAR; AI; edge computing; sensor fusion.

I. Introduction

As urban areas grow, the need to monitor road conditions efficiently becomes crucial for keeping infrastructure intact and promoting road safety. The conventional methods of inspecting roads are laborious, time-consuming, and frequently fall short of providing the accuracy required for proactive repairs. However, recent progress in sensor technology, artificial intelligence, and data integration presents fresh opportunities for transforming road condition monitoring.

Our approach overcomes this challenge by basing AI inference solely on standard Red-Green-Blue (RGB) images and then projecting the 2D detection information into the 3D domain provided by the LiDAR. This is achieved through a meticulous camera-LiDAR calibration process, as shown in Figure 1.

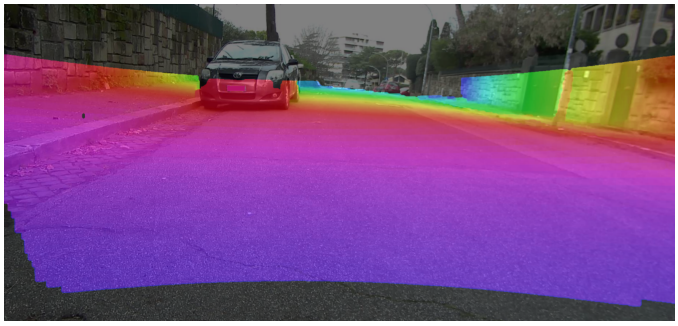


Figure 1. Camera-LiDAR Registration.

This strategy allows us to leverage large, pre-existing public datasets for training, significantly reducing development time and cost. As an extension of our preliminary work [1], this paper details the complete system that utilizes LiDAR technology along with RGB imaging and GNSS/INS data within a Robot Operating System (ROS) based framework to identify and map road surface issues efficiently. We provide a deep dive into the system's architecture, the full AI pipeline, and a comprehensive report on its real-world performance, including a transparent discussion of the operational challenges we overcame.

From an economic standpoint, the system's adaptability to city vehicles, including public transport, could potentially transform routine operations into continuous, cost-effective road monitoring. Combining this distributed sensing with on-the-ground human supervision, such as cleaning personnel, creates a hybrid model that optimizes resource use and enhances data accuracy, leading to efficient urban road maintenance.

The paper is structured as follows. Section II discusses the state of the art. Section III outlines the specific advancements over our previous conference paper. Section IV describes the system's hardware and software architecture. Section V details the AI pipeline. Section VI presents the experimental validation results, and Section VII concludes the paper.

II. Related Work

Over the past few years, many approaches have been explored for automated road inspection. Some methods rely on inertial data [2], while others utilize pure machine learning and computer vision techniques [3][4]. More sophisticated approaches exploit deep learning models [5] or combine vision and depth sensing together with spatial AI [6][7].

The technologies that have been tested for depth estimation are based on stereoscopy, Red-Green-Blue-Depth (RGB-D) cameras, and LiDAR. However, each has its own disadvantages: stereoscopy generally does not work well with feature-poor surfaces. RGB-D cameras based on Time of Flight (ToF) technology, while achieving good accuracy, drop their performance in outdoor environments and are limited to a range of a few meters. Conversely, LiDAR provides the most long-range and accurate measurements but at the expense of lower point density and the need for an additional imaging system to obtain the scene picture. Furthermore, approaches using RGB-D images as input for AI detection models, while achieving good performance due to depth information, are strongly affected by the context, sensor position, and framing of the training data. Therefore, they require the acquisition of huge amounts of images from every possible angle and distance in order to replicate all possible setups.

III. Extension of Previous Work

This journal article represents a substantial extension of the preliminary research presented in our conference paper [1]. While the original work introduced the concept of the multi-sensor fusion architecture, this paper incorporates significant technical advancements and a more rigorous validation methodology.

Firstly, the AI pipeline has been refined. In [1], the focus was primarily on detection feasibility. In this work, we present a consolidated dual-model approach (YOLOv8-seg and SegFormer) with a dedicated section on the dataset curation process, including the integration of negative examples for manholes to reduce false positives.

Secondly, the experimental validation has been vastly expanded. The previous work relied on limited datasets. Here, we present results from an extensive on-site campaign covering approximately 70 km of urban roads in Rome. This includes a new quantitative analysis of telemetry accuracy (area, volume, depth) and geolocation precision compared to ground truth.

Thirdly, we include a detailed discussion on the iterative tuning process required to handle real-world environmental challenges, such as shadows and motion blur, which were not addressed in the initial study. This comparison underscores the transition from a proof-of-concept to a field-validated prototype.

IV. System Architecture

The system was engineered as a modular, vehicle-mounted unit designed for robust data acquisition in dynamic urban environments. Its architecture integrates carefully selected hardware components with a sophisticated software pipeline built on ROS to ensure interoperability and scalability.

A. Hardware Configuration

The hardware setup was chosen to balance high-performance data acquisition with resilience to on-road conditions and suitability for edge computing. An NVIDIA Jetson AGX Orin 64GB module serves as the central computing unit, providing the necessary power for real-time AI inference. For 3D perception, an HESAI Technology AT128 Hybrid Solid-State LiDAR was selected for its optimal Field of View (FOV) and mechanical resilience. Visual information is captured by a 4K global shutter camera, a critical choice made to eliminate motion blur. A Microstrain 3DM-GQ7 module provides precise geolocalization and orientation by fusing data from a dual-antenna GNSS receiver and a 9-axis IMU. The entire system is supported by a 4G/LTE router, a network switch, and a dedicated power management system. The overall setup is shown in Figure 2.

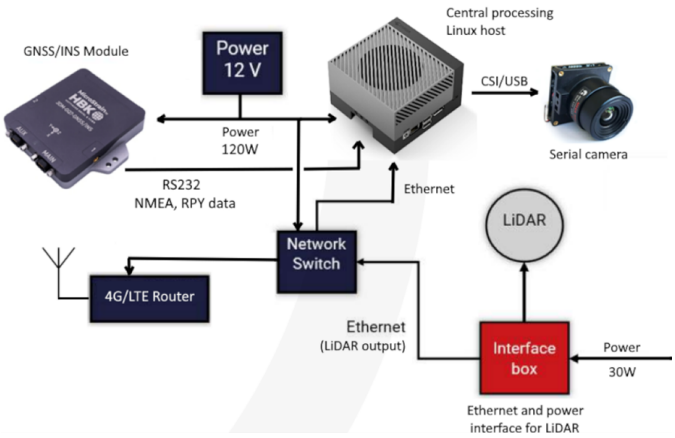


Figure 2. Hardware setup.

B. Software Architecture

The software was built entirely on ROS 2, a flexible framework for communication and data processing. The architecture functions as a pipeline that transforms raw sensor data into actionable insights. The process begins with dedicated driver nodes that interface with each sensor and publish data onto specific ROS topics. A core AI Inference Node, developed for this project, subscribes to these topics and uses an internal synchronizer to create a coherent snapshot of the environment from different sensor inputs. This synchronized data is then fed into the AI models. The resulting output is packaged into Safetensors files for later use. A separate Post-Processing and Reporting Node operates independently on these files. This agent performs the final data fusion and analysis, projecting 2D detections onto the 3D LiDAR point cloud, calculating the geometric properties of each defect, and assigning precise geographic coordinates. Finally, it formats the data into a JSON payload for transmission to a map-based web portal. This decoupled architecture makes the process resilient to network connectivity issues.

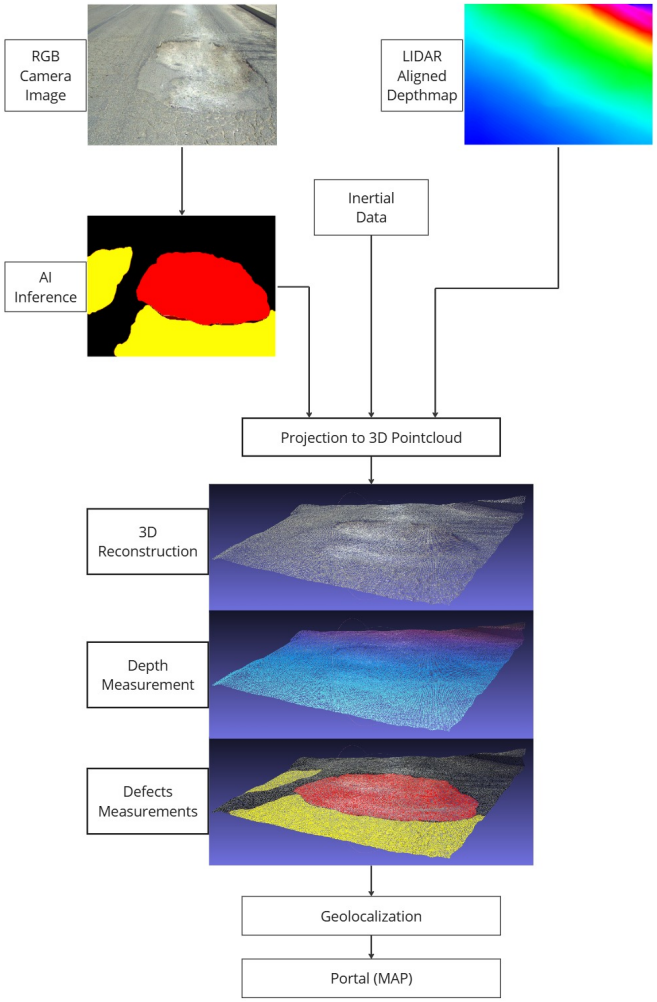


Figure 3. Software architecture components in ROS framework.

V. AI Models

The system's ability to accurately identify road defects depends on an AI pipeline composed of two distinct deep learning models. The development process involved careful model selection, extensive dataset preparation, rigorous training, and detailed validation.

A. Model Selection

Two primary architectures were chosen for their proven performance. For defect detection, a You Only Look Once (YOLO) model [8] version "v8small-seg" was selected for its powerful instance segmentation capabilities. This task extends beyond simple object detection by not only providing a bounding box for each detected object but also generating a pixel-perfect mask that outlines its exact shape. This capability is crucial for our application, as the generated masks are later projected into the 3D domain to enable precise geometric measurements of defects, such as their area and volume. The "s" variant (YOLOv8s-seg) was specifically chosen as it offers an excellent trade-off between accuracy and computational efficiency, making it highly suitable for real-time processing on our edge computing platform.

In parallel, a SegFormer variant B1 model [9] is used for scene understanding. This model is based on a Vision Transformer (ViT) architecture, which, unlike traditional CNNs, excels at capturing long-range dependencies and global context within an image. This makes SegFormer particularly effective for semantic segmentation, the task of assigning a class label to every pixel. Its primary role in our pipeline is to generate a highly accurate and reliable "road mask". This mask serves as a critical contextual filter: by intersecting the defect detections from YOLO with this road mask, we can effectively eliminate false positives that may occur on sidewalks, vegetation, or other non-road surfaces, thereby significantly increasing the overall reliability of the system.

B. Dataset Preparation and Training

The performance of the AI models relies on carefully prepared training data. For the scene understanding model, a transfer learning approach was used, employing a SegFormerB1 model pre-trained on the well-known Cityscapes dataset [10]. The dense annotations of urban scenes in this large-scale dataset enabled high-performance road segmentation without the need to create a new dataset from scratch.

For the defect detection model, however, an initial analysis of existing public datasets revealed that none fully met the project's requirements in terms of camera perspective, labeling quality, and class definitions. Consequently, a significant effort was dedicated to creating a custom, high-quality dataset. The process began by curating a base set of images from the public Road Damage Detection (RDD) dataset [11], selecting only those from geographical regions with road conditions and perspectives relevant to the target operational environment.

This base set was strategically composed to include a substantial number of images without any defects to train the model to minimize false positives on well-maintained road surfaces. A critical challenge identified was the underrepresentation of certain scenarios, particularly images containing both potholes and manholes, which often led to misclassifications. To address this, the dataset was enriched with hundreds of additional images sourced from various other public repositories, specifically chosen to increase the variety of potholes and provide negative examples of manholes.

The final curated dataset consisted of 6,504 images, split into training (5,199), validation (652), and test (653) sets. A meticulous manual re-labeling process was undertaken with the help of Segment Anything Model (SAM) [12] to create precise instance segmentation masks for two target classes: pothole and crack, which consolidated various types of fissures like alligator and linear cracks into a single category. To further enhance the model's robustness and its ability to generalize, the training set was expanded through extensive data augmentation. Techniques such as flipping, rotation, and adjustments to saturation, brightness, and noise were applied, resulting in a final training dataset of 25,995 images.

The training process itself employed a fine-tuning strategy, starting from the Common Objects in Context (COCO) pre-trained model. This approach leverages the generalized features learned on a large-scale dataset and adapts them to the specific task of defect detection. The default training configurations were used, which include additional data augmentation techniques like mosaicing to improve the model's performance on objects at various scales. The training was configured to run for 100 epochs with a standardized input image size of 640x640 pixels, using the custom dataset described above.

C. Model Validation and Performance Metrics

After training, the models were rigorously evaluated on their respective test sets, to provide an unbiased assessment of their generalization capabilities. This validation involved both a quantitative analysis through standard computer vision metrics and a qualitative visual inspection of the model's predictions. The quantitative evaluation is based on the confusion matrix, which categorizes predictions into True Positives (TP), False Positives (FP), and False Negatives (FN). Key metrics include Precision, which measures the model's ability to avoid false positives, and Recall, which measures the model's ability to find all relevant instances in an image. Additionally, also the F1-score is taken into account, providing the balance between Precision and Recall in one formula:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{1}$$

where P is the Precision and R is the Recall value.

For tasks involving spatial localization, such as segmentation and detection, accuracy is quantified by the Intersection over Union (IoU). This metric measures the overlap between the predicted region (mask or bounding box) and the ground-truth region, providing a score for spatial accuracy. From this, the Average Precision (AP) is calculated for each class by averaging the precision values over the Precision-Recall curve. The primary summary metric for object detection models is the mean Average Precision (mAP), which is the mean of the AP values across all classes and, often, across a range of IoU thresholds, e.g., mAP@0.5:0.95. The mAP is calculated with the following equation:

$$\text{mAP} = \frac{1}{N_c} \sum_{c=1}^{N_c} \frac{1}{N_{\text{IoU}}} \sum_{i=1}^{N_{\text{IoU}}} \text{AP}_c^{(i)} \tag{2}$$

where N_c is the total number of classes, N_{IoU} is the number of IoU thresholds, and $\text{AP}_c^{(i)}$ is the average precision for class c at IoU threshold i.

The YOLOv8s-seg defect detection model achieved an mAP of 0.564. While the performance for the pothole class was strong, the instance segmentation of cracks proved more challenging due to their ambiguous and continuous nature, making it difficult for the model to distinguish separate instances. However, when evaluating the crack class from a semantic segmentation perspective (merging all predicted crack masks), the model's ability to correctly identify crack pixels versus background was excellent, confirming its effectiveness for the project's use case. The F1-Confidence curve indicated an optimal confidence threshold of 0.343, at which the model reached a balanced F1-score of 0.57.

The SegFormerB1 model demonstrated outstanding performance for road segmentation. On the Cityscapes test set, it achieved an F1-score of 0.98, specifically for the main road class, with a correct pixel classification rate of 99%. On the other hand, semantic segmentation models are often evaluated with meanIoU metrics, calculated with the following formulas:

$$\text{IoU}_c = \frac{|A_c \cap B_c|}{|A_c \cup B_c|} \tag{3}$$

$$\text{meanIoU} = \frac{1}{N_c} \sum_{c=1}^{N_c} \text{IoU}_c \tag{4}$$

where A_c is the predicted segmentation for class c, B_c is the ground truth segmentation for class c, and N_c is the number of classes.

The resulting meanIoU was actually 0.43, not as excellent as other metrics, but this is expected as the metric heavily penalizes minor shape deviations, which are common in complex road scenes. For the primary task of creating a reliable road mask, the performance was deemed excellent. Following validation, both models were optimized using NVIDIA TensorRT with FP16 precision, which more than doubled their inference speed on the edge device with negligible impact on accuracy. A summary of the resulting metrics is shown in Table I. A qualitative visual analysis further confirmed the two models' performance and robustness across various scenarios as shown in Figure 4.

TABLE I. SUMMARY OF AI MODEL PERFORMANCE.

Model	Main Task	Key Metric	Value
Yolov8s-seg	Road Defect Detection	mAP@0.5 (all classes)	0.564
		F1-Score (optimal)	0.57
SegFormerB1	Road Segmentation	F1-Score ("Road" class)	$\approx$ 0.98
		MeanIOU ("Road" class)	0.43

VI. EXPERIMENTAL SYSTEM VALIDATION

The system's validation followed a two-stage process. First, controlled laboratory tests were conducted to calibrate the sensors and benchmark core functionalities. Then, extensive on-site tests were run to evaluate its performance and robustness in real-world scenarios.

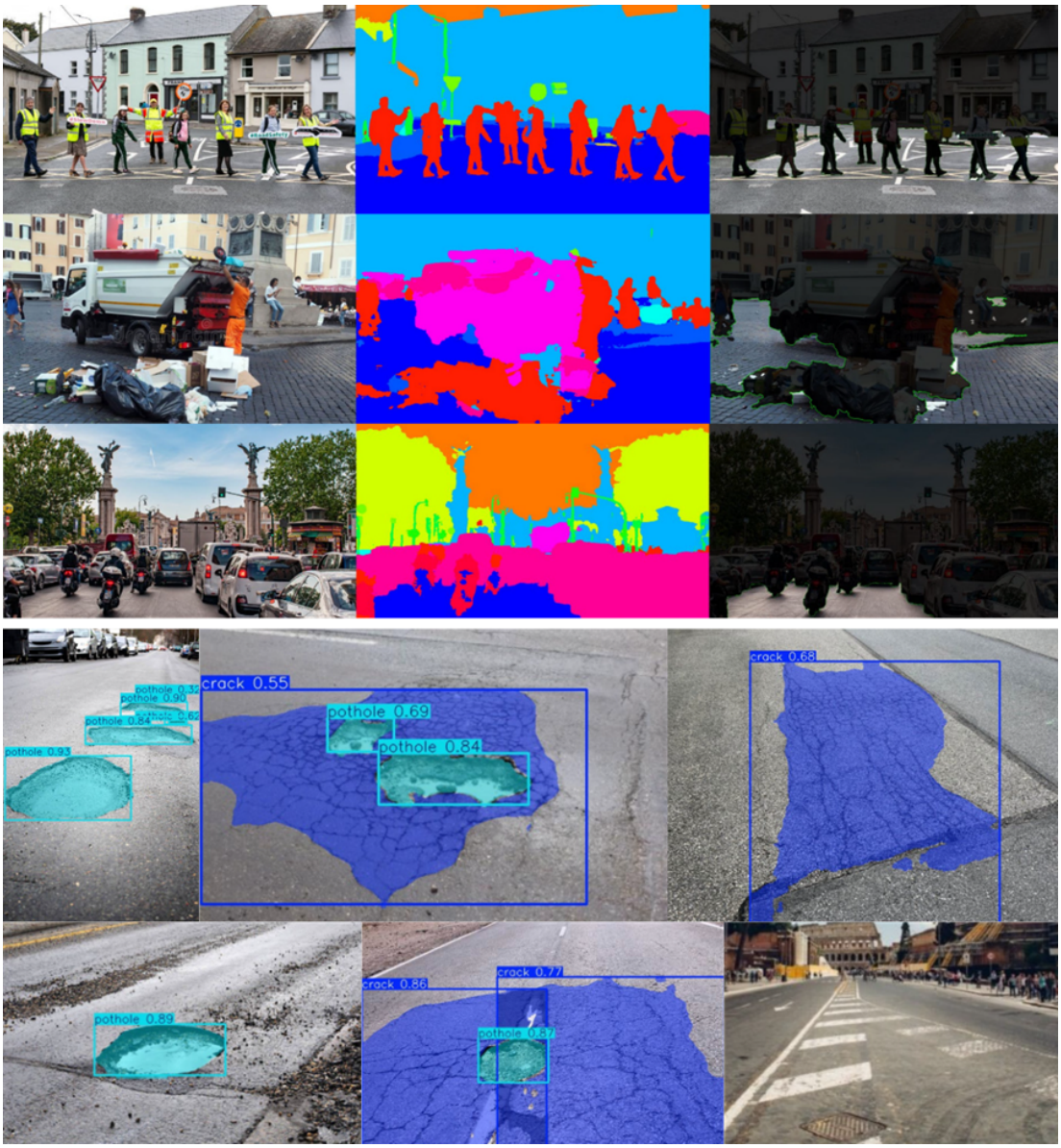


Figure 4. Qualitative inspection of model outputs. The top rows illustrate the SegFormerB1 model's road segmentation, displaying the input image, the complete segmentation map colored by CityScapes class, and the activation map for the "road" class. The bottom rows present a mosaic of detections from the YoloV8s-seg model, which identifies cracks (blue) and potholes (cyan) using bounding boxes and masks, each annotated with a confidence score.

A. Laboratory Calibration and Testing

Before on-site deployment, the integrated prototype was subjected to a series of crucial tests in a controlled laboratory environment. This preparatory phase was fundamental to ensuring the system's reliability and accuracy.

The most critical step was the multi-sensor calibration. Since the system relies on fusing data from a 2D camera and a 3D LiDAR, it was essential to precisely determine the geometric relationship between them. This process was divided into two parts. First, an intrinsic camera calibration was performed using a standard chessboard pattern viewed from multiple angles, as shown in Figure 5.

This allowed us to calculate the camera's internal parameters, i.e., the K matrix shown in Table II, and to generate correction maps to remove lens distortion. As shown in Figure 6, this undistortion process transforms the raw, warped image into a geometrically accurate one, which is a prerequisite for any precise measurement.

TABLE II. INTRINSIC MATRIX (K).

$$K = \begin{bmatrix} 304.6712 & 0.0 & 313.5861 \\ 0.0 & 380.5664 & 165.4646 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}$$

Next, an extrinsic LiDAR-camera calibration was conducted. This procedure establishes the rigid transformation, i.e., rotation and translation, between the LiDAR's and the

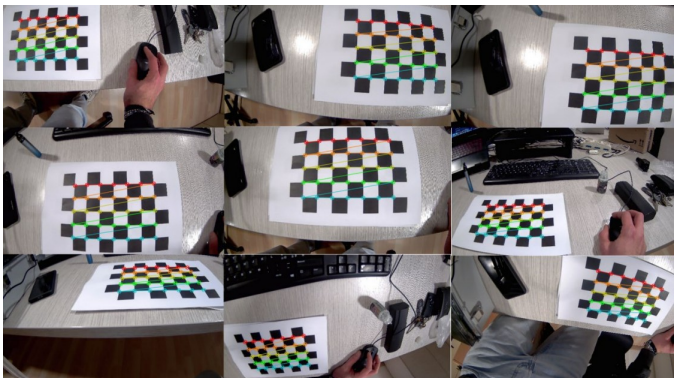


Figure 5. Intrinsics calibration process: using multiple images of chessboard from different points of view to calculate intrinsics matrix.

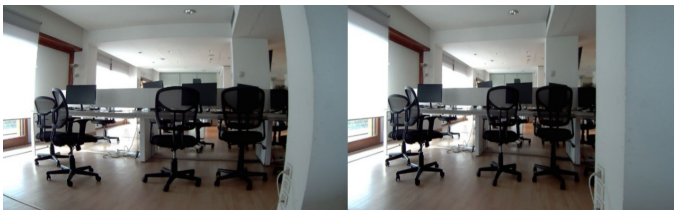


Figure 6. Undistortion process, after intrinsics calibration. Left: original image. Right: Undistorted image.

camera's coordinate systems. By manually identifying a set of corresponding points in both the 3D LiDAR point cloud and the 2D camera image, we used a Perspective-n-Point (PnP) algorithm to compute the 4x4 roto-translation matrix, shown in Table III. The accuracy of this calibration was verified by reprojecting the 3D LiDAR points onto the 2D image using the calculated matrix; the near-perfect alignment confirmed the success of the calibration. This matrix is the key that enables the accurate fusion of data from the two sensors. Final results of reprojection are shown in Figure 7.

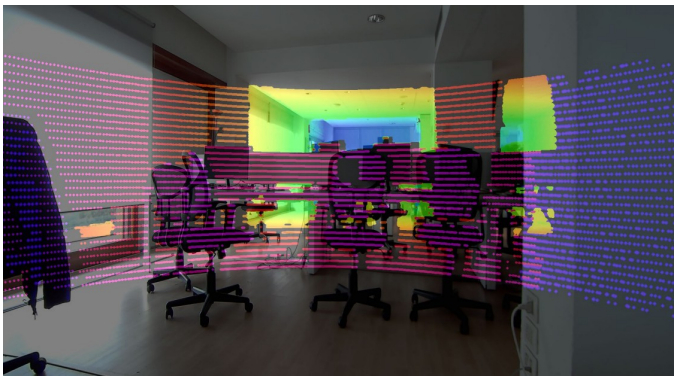


Figure 7. Point cloud reprojection onto RGB image, after intrinsics and estrinsics calibration.

With the sensors calibrated, we proceeded to benchmark the system's core functionalities. Throughput tests on the ROS AI node were conducted to measure its computational performance. The system demonstrated a stable processing average of 11.12 Frames Per Second (FPS), confirming its capability for real-time processing at typical urban driving speeds.

TABLE III. ESTRINSINCS ROTO-TRANSLATION MATRIX (RT).

$$RT = \begin{bmatrix} 0.998 & -0.008 & -0.059 & -0.111 \\ 0.008 & 1.000 & 0.005 & 0.014 \\ 0.059 & -0.005 & 0.998 & 0.093 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}$$

The system's telemetry accuracy was evaluated using targets of known size - specifically, two manholes of different shapes - to simulate road defects. The results were highly encouraging, showing a low average relative error of just 6% for surface area measurements. The more challenging depth and volume estimations also yielded respectable average relative errors of 24% and 19%, respectively, demonstrating a good approximation capability.

Finally, static geolocation tests were performed to assess positioning accuracy. The system reported the coordinates of known points, which were then compared against high-precision ground truth data obtained from a topographic survey. The tests revealed a mean horizontal error of 2.13 meters, an accuracy level that is well within acceptable limits for the primary goal of dispatching maintenance crews to the correct location. A summary of the resulting metrics is shown in Table IV.

TABLE IV. SUMMARY OF THE PROTOTYPE'S MEASUREMENT AND POSITIONING PERFORMANCE.

Category	Metric	Calculated Value
Defect Dimensional Estimation (Telemetry)		
Mean Relative Error	Area	6%
Mean Relative Error	Depth	24%
Mean Relative Error	Volume	19%
Geographic Defect Localization		
Mean Error	Horizontal	2.131 meters
Root Mean Squared Error	Horizontal RMSE	2.780 meters

B. On-Site Testing

The most comprehensive and conclusive validation of the system was an extensive on-site testing campaign. This involved deploying the fully calibrated prototype on a vehicle and conducting multiple data acquisition sessions across approximately 70 km of public roads in Rome, Italy. The system has been positioned on a vehicle, pointing forward as shown in Figure 9. The routes were strategically selected to cover a wide spectrum of real-world conditions, including different road types (from high-speed arteries to narrow residential streets), varying traffic densities, and diverse lighting environments. This campaign was structured as an iterative process of testing, analysis, and refinement, allowing to systematically address challenges encountered in the field.

A primary challenge identified during the initial test runs was the misclassification of manholes. The AI model fre-

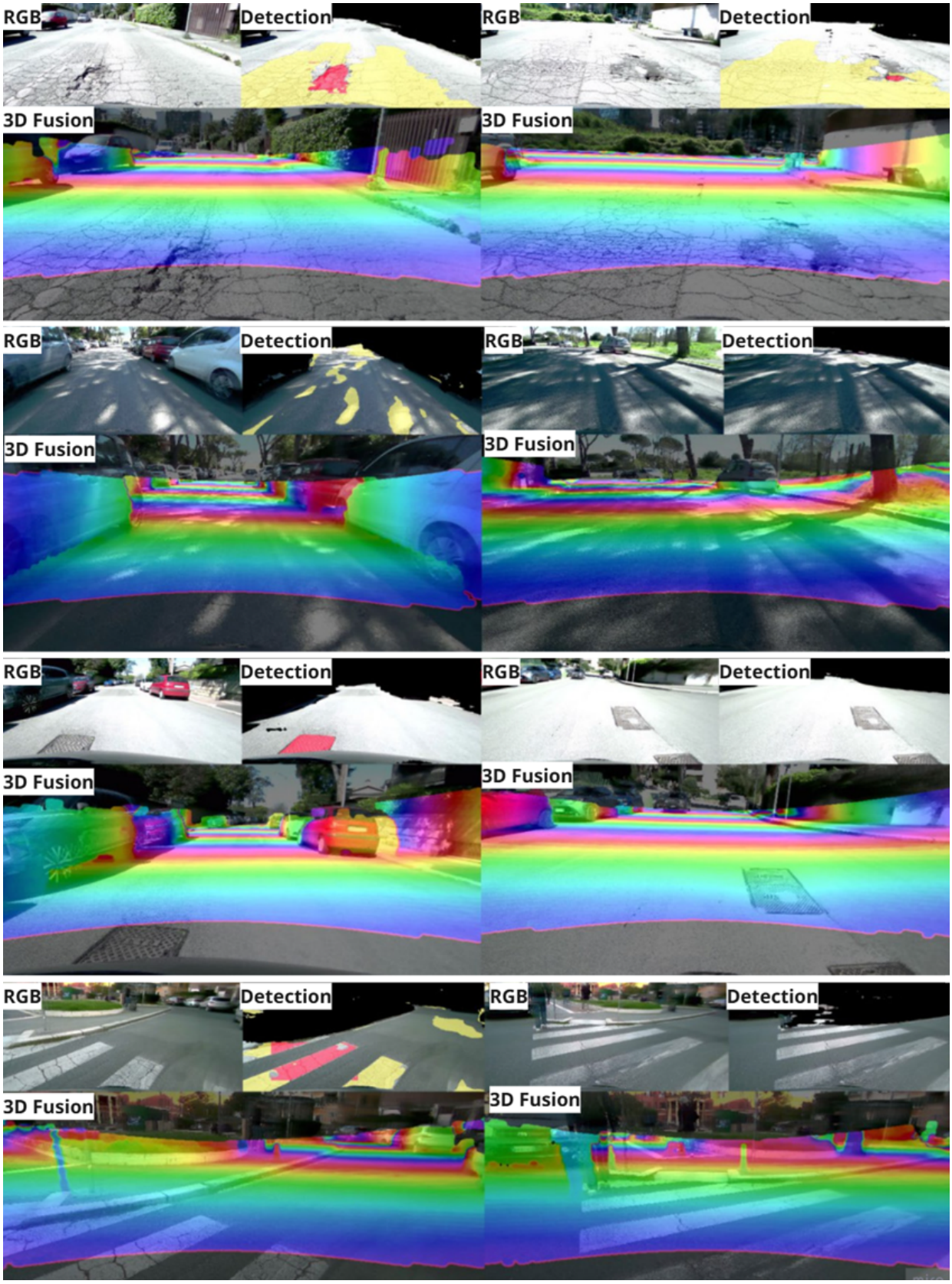


Figure 8. Iterative tuning of on-site tests: each row represents different situations. On the left: the system output before correction. On the right: output after tuning. The main issues were related to: false positives related to manholes, sharp shadows producing fake crack output, dark environment and road marks sensitivity.

Figure 9. A picture of the system prototype.

Figure 10. A screenshot of the web application for the visualization of results.

quently identified manholes, particularly those slightly recessed or with damaged edges, as "potholes". Through a careful analysis of these false positives, it has been determined that the confidence scores assigned to manholes were consistently lower than those for genuine potholes. Based on this observation, we iteratively adjusted the detection threshold, raising the minimum score threshold value for potholes to 0.30. This seemingly simple change proved highly effective, making the model more selective and drastically reducing manhole-related false positives without compromising its sensitivity to actual defects.

Another significant issue was caused by environmental factors, specifically the strong, hard-edged shadows cast by buildings and trees on sunny days. The high contrast along these shadow lines was often misinterpreted by the model as "cracks". This led to another round of data-driven tuning. After experimenting with different values, the confidence threshold for the minimum score threshold value for cracks was also set to 0.30. This found an optimal balance, successfully eliminating the vast majority of shadow-induced artifacts while still reliably detecting significant cracks, and also helped in correctly ignoring most worn-out road markings. Examples of challenges overcome through iterative tuning are shown in Figure 8.

The campaign culminated in a final, long-duration test session of approximately 35 km, which served as a validation run for the fully tuned system. This test confirmed the system's operational stability and thermal resistance over an extended period. Throughout this iterative process, the web portal as shown in Figure 10, proved to be a valid tool, allowing for rapid qualitative inspection of results and enabling the data-driven refinements that led to a robust and reliable final configuration for automated road condition assessment.

VII. Conclusion and Future Work

This paper has presented an extended and in-depth analysis of a multi-sensor system for automated road defect assessment, building upon previous work [1]. By detailing the system's architecture, its AI pipeline, and the results of extensive validation, this work has demonstrated a robust and viable solution for enhancing Smart City infrastructure management. The rigorous testing campaign confirmed the system's high performance. The AI-driven pipeline achieves reliable detection of critical defects suitable for practical applications. Furthermore, the fusion of LiDAR and camera data enables quantitative measurements and geolocalization with a precision that provides actionable data for maintenance planning. The iterative, data-driven tuning of the AI models was essential for overcoming real-world challenges and achieving a stable final configuration. The deployment of the entire processing pipeline on an edge computing device ensures real-time capabilities and operational autonomy. The final system represents a significant step forward from conventional inspection methods, offering a scalable, cost-effective, and data-rich approach to proactive road maintenance. Future work will specifically focus on improving the segmentation performance for complex crack patterns by exploring advanced model architectures and expanding the dataset with more diverse crack topologies, alongside further refining depth and volume estimation algorithms. Ultimately, this work validates the power of integrating advanced sensing and artificial intelligence to create smarter and more efficient urban environments.

Acknowledgement

This project has been co-financed by the European Union through the PR FESR 2021–2027 RSI program of Regione Lazio, managed by LazioInnova. The authors would like to thank the European Union and Regione Lazio for their support in enabling this research. Additionally, we extend our gratitude to all partners and collaborators who contributed to the successful implementation and validation of the proposed system.

References

[1] G. Nardini et al., "Smart city road maintenance: A lidar and ai-driven approach for detecting and mapping road defects", in *The Fourteenth International Conference on Smart Cities, Systems, Devices and Technologies (SMART 2025)*, Valencia, Spain, Apr. 2025.

[2] S. Girisan, T. V. Sreelakshmi, N. V. Swetha, V. Suresh, and K. M. Vipin, "Pothole detection based on accelerometer method", *International Journal of Research in Engineering, Science and Management*, vol. 3, no. 5, May 2020.

[3] A. Ahmadi, S. Khalesi, and M. R. Bagheri, "Automatic road crack detection and classification using image processing techniques, machine learning and integrated models in urban areas: A novel image binarization technique", *Journal of Industrial and Systems Engineering*, vol. 11, pp. 85–97, 2018.

[4] K. Li, B. Wang, Y. Tian, and Z. Qi, "Fast and accurate road crack detection based on adaptive cost-sensitive loss function", *IEEE Transactions on Cybernetics*, vol. 53, no. 2, pp. 1051–1062, Feb. 2023, ISSN: 2168-2275.

[5] Y. Li, C. Yin, Y. Lei, J. Zhang, and Y. Yan, "Rdd-yolo: Road damage detection algorithm based on improved you only look once version 8", *Applied Sciences*, vol. 14, p. 3360, Apr. 2024.

[6] E. M. Thompson et al., "Shrec 2022: Pothole and crack detection in the road pavement using images and rgb-d data", in *SHREC2022 3D Shape Retrieval Challenge*, 2022.

[7] A. Talha, M. Karasneh, D. Manasreh, A. Oide, and M. Nazzal, "A lidar-camera fusion approach for automated detection and assessment of potholes using an autonomous vehicle platform", *Innovative Infrastructure Solutions*, vol. 8, Sep. 2023.

[8] J. Zeng, H. Ouyang, M. Liu, L. U. Leng, and X. Fu, "Multi-scale yolact for instance segmentation", *Journal of King Saud University - Computer and Information Sciences*, vol. 34, Oct. 2022.

[9] E. Xie et al., "Segformer: Simple and efficient design for semantic segmentation with transformers", in *Neural Information Processing Systems (NeurIPS)*, 2021.

[10] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223.

[11] D. Arya, H. Maeda, S. K. Ghosh, D. Toshniwal, and Y. Sekimoto, "Rdd2022: A multi-national image dataset for automatic road damage detection", in *Crowdsensing-based Road Damage Detection Challenge (CRDDC)*, 2022.

[12] A. Kirillov et al., "Segment anything", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026.

Identifying Factors that Increase the Risk of Demotivation in Scientific Computing Courses Using Monte Carlo Methods

Isaac Caicedo-Castro*†‡ Rubby Castro-Púche†§ Oswaldo Vélez-Langs*‡
*Socrates Research Team
†Research Team: Development, Education, and Healthcare
‡Faculty of Engineering
§Faculty of Education and Human Science
University of Córdoba
Carrera 6 No. 76-305, 230002, Montería, Colombia
e-mail: {isacaic | rubycastro | oswaldovelez}@correo.unicordoba.edu.co

***Abstract*—This study aims to identify the factors associated with the risk of demotivation in scientific computing courses. To achieve this, we modeled the functional relationship between student motivation and influencing factors using supervised machine learning, particularly Bayesian regression. This relationship was then incorporated into a Monte Carlo simulation to generate a wide range of scenarios, allowing us to estimate both absolute and relative risks of demotivation for each factor. In conclusion, the results reveal that the strongest predictors of increased demotivation risk are low levels of student satisfaction and enjoyment, followed by insufficient encouragement of independent study and limited access to up-to-date equipment, among other factors.**

Keywords-higher education; motivation; Bayesian regression; Monte Carlo methods.

I. Introduction

In this study, we identify the factors associated with an increased risk of demotivation among students enrolled in scientific courses. To this end, we applied regression and Monte Carlo statistical methods to estimate students' motivation levels across a wide input space defined by multiple motivation factors. These estimated motivation levels were then used to compute both the absolute and relative risks of demotivation.

Identifying these factors is essential for designing effective policies and implementing strategies to prevent or mitigate the risk of demotivation. Scientific courses, such as numerical methods, are particularly challenging because they combine mathematics, computer programming, and scientific domains (e.g., physics, chemistry, biology), each of which is already difficult on its own. Additional difficulties arise from the abstract mathematical concepts underlying scientific computing, as well as from students' struggles to understand how these methods can be applied to solve real-world problems across diverse scientific fields.

Reducing the risk of demotivation is critical, as demotivated students often lack the willingness or drive to complete assignments, prepare for examinations, and engage with learning activities. In the context of scientific computing courses, sustaining student motivation is particularly challenging.

For this study, we collected data from a community sample of students enrolled in the Systems Engineering bachelor's program at the University of Córdoba, Colombia. This dataset was used to estimate a function that predicts students' motivation based on the values of influencing factors

Using this functional dependency, we applied the Monte Carlo method to simulate a broader range of values for the influencing factors than those available in the dataset. The goal of this simulation was to estimate the risk of demotivation. In this context, simulation provides an appropriate alternative to avoid unethical experiments in which students would be exposed to stressful or unfavorable scenarios in order to directly observe demotivation risk.

The results of the Monte Carlo simulation indicate that targeted interventions are needed to design courses that foster student satisfaction and enjoyment, as both factors are strongly associated with a high risk of demotivation in scientific computing courses. Interventions are also necessary in prerequisite mathematics courses, since the greatest risk of demotivation was linked to students' experiences in prior mathematics coursework. Enhancing satisfaction and enjoyment in these foundational courses may therefore reduce the overall risk of demotivation in subsequent scientific computing studies.

Additional factors associated with the risk of demotivation include:

i) Access to up-to-date equipment to support scientific computing courses.
ii) Encouragement for independent study, cooperation, and collaborative coursework.
iii) Student focus and engagement in course activities.
iv) Student beliefs regarding the usefulness of the course and mathematics in general for their future professional life, their self-perceived ability to learn mathematics and solve mathematics-related problems, and their perception of the importance of hard work for successfully completing the course.

By adopting a Bayesian regression model, our study achieved a modest improvement in predictive performance, increasing the coefficient of determination from 0.37 reported in prior research [1] to 0.38. Moreover, unlike aforementioned previous research, we explicitly computed and analyzed both the relative and absolute risks associated with each factor linked to demotivation, thereby providing a more nuanced

and actionable assessment of how these factors contribute to students' motivational outcomes.

Absolute risk represents the simulated probability of demotivation and reflects the practical impact of each factor on the student population. In contrast, relative risk measures the strength of association by comparing the probability of demotivation between exposed and non-exposed groups, thereby indicating the extent to which a given factor increases or decreases risk relative to a baseline condition. Taken together, these metrics enable the identification of factors that are not only statistically associated with demotivation but also substantively meaningful in practical terms.

The remainder of this paper is outlined as follows: we discuss the literature review in Section II, and present the methodology adopted in this research in Section III. We report and analyze the results in Sections IV and V. Finally, we conclude the paper in Section VI and propose directions for further research.

II. Prior Research

Learning scientific computing is challenging because students must integrate knowledge of mathematics, computer programming, and the sciences (e.g., physics, chemistry). Mathematics is essential for understanding how numerical methods work, while computer programming is required to implement them. Moreover, solving real-world engineering problems demands a solid grounding in science to understand the problem context and to apply numerical methods effectively.

This challenge has motivated research aimed at predicting which students are at risk of failing scientific computing courses based on their performance in prerequisite subjects [2][3]. Recent studies have even explored quantum machine learning approaches to address this problem [4]. Furthermore, prediction accuracy has been improved by adopting alternative representations of the independent variables, considering only student performance in prerequisite mathematics courses (i.e., linear algebra, differential, integral, and vector calculus) [5].

The findings in [5] highlight the importance of students' mathematics background for success in scientific computing courses. However, learning mathematics is itself a challenging task. Consequently, identifying the factors that influence mathematics learning has been a subject of extensive research, ranging from basic education [6][7][8][9] to higher education [10][11][12][13][14], and even at the doctoral level [15]. Scientific computing, essentially an applied mathematics discipline, encompasses numerical methods and heuristics for solving mathematical problems in science and engineering that cannot be addressed analytically.

In Colombia, studies have explored the process of knowledge construction among college students in the context of algebra courses within engineering curricula [11]. Previous research has primarily focused on students' commitment, satisfaction, and the challenges they face in learning mathematics at the college level.

Students' motivation for learning scientific computing has been investigated using machine learning, particularly regression, and the Monte Carlo method to estimate the probability that a student reaches one of ten motivation levels. The study was conducted with 117 students enrolled in scientific computing courses within the undergraduate Systems Engineering program at the University of Córdoba in Colombia. The results revealed that students are most likely to achieve moderate motivation levels, although effective policies and strategies could increase the probability of attaining higher levels of motivation [1].

In this paper, we estimate the functional dependency between independent and dependent variables using linear regression fitted with the No-U-Turn Sampler (NUTS) [16], a Hamiltonian Monte Carlo method. This differs from the previous study [1], which adopted Ridge Regression (cf. [17] for details). We then use the regression function to explore a broader space of independent variables in order to calculate the probability that a student reaches a specific motivation level, as in [1]. In addition, we estimate the absolute and relative risk of demotivation associated with the independent variables, an analysis that, to our knowledge, has not been conducted in prior research.

III. Research Methodology

We adopted a quantitative approach in which the factors assumed influence the students' motivation to study scientific computing courses are treated as independent variables, while the student's motivation level serves as the dependent variable (aka, target variable). Our goal is to estimate the functional dependency between the independent and dependent variables, i.e., to identify the function $g : \mathcal{X} \to \mathcal{Y}$ that maps, for the ith student, the independent variables represented by the D-dimensional vector $x_i \in \mathcal{X} \subset \mathbb{R}^D$ to the dependent variable $y_i \in \mathcal{Y} \subset \mathbb{R}$. Herein, we consider $D = 15$ independent variables.

Thus, the function $g(x_i)$ predicts motivation level of the ith student given their influential factors x_i. Henceforth, the vector x_i shall be referred to as the input variables or input vector, since its component serve as input to the function g. We consider the same input variables utilized in [1], as listed in Table I. Some of these variables are also used in [13][14]. Each factor is quantified on a scale from 1 to 5 and then rescaled to the interval $[0, 1]$. For instance, if the ith student perceives that the university provides up-to-date equipment and assigns this factor a value of 5, the corresponding input variable becomes $x_{i,5} = 1$.

On the other hand, the dependent variable is measured on a discrete scale from 1 to 10, where higher values indicate greater student motivation. Hereafter, y_i is referred to as the output variable, since the function g approximates it (i.e., $g(x_i) \approx y_i$).

We used the same dataset collected in 2024 by Caicedo-Castro et al. [1], which contains 117 examples obtained from a survey of students enrolled in scientific computing courses, specifically Numerical Methods and Nonlinear Programming. The identities of the students were anonymized. The input variables were selected using an F-test: if the null hypothesis of no linear relationship between a given input variable and the output variable was rejected (i.e., p-value $< 5 \times 10^{-2}$), the variable will be included for fitting the regression model.

Following this criterion, a total of 15 input variables were retained, as listed in Table I.

The histogram, shown in Figure 1, illustrates that the maximum motivation level was chosen by most of the students, namely, 48 out of 117 students (see Table II).

Given the dataset described above, we perform a Bayesian linear regression to estimate the function g that models the relationship between the predictors and student motivation. In this framework, the regression parameters are treated as random variables with prior probability distributions. The model is defined as:

$$\hat{y}_i = \beta^T x_i + \beta_0 + \epsilon_i \tag{1}$$

where the real-valued D-dimensional vector $\beta \in \mathbb{R}^D$ and the real number β_0 are the parameters or weights (aka., coefficients) of the function g. Besides, $\hat{y}_i$ denotes the predicted motivation level for student ith $\hat{y}_i \approx y_i$, $\epsilon_i \sim \mathcal{N}(0, \sigma)$ represents the error term, and σ is the standard deviation of the residuals. The prior distribution are specified as follows: $\sigma \sim \mathcal{N}(0, 1)$, and the weights $\beta_j \sim \mathcal{N}(0, 1)$, for $j = 0, \ldots, D$. Consequently, the likelihood of the observations is given by:

$$y_i \sim \mathcal{N}(\beta^T x_i + \beta_0, \sigma) \tag{2}$$

The regression model was implemented in PyMC[18], which performs Bayesian posterior sampling using the No-U-Turn Sampler (NUTS). NUTS is an extension of Hamiltonian Monte Carlo (HMC) that adaptively determines the number of leapfrog steps L, thereby avoiding the need to specify this tuning parameter manually. This is advantageous because choosing L too small induces random-walk behavior, whereas an excessively large L results in unnecessary computational overhead (cf. [16] for details). The sampler was run with 14 parallel chains, each drawing 3500 samples, and a target acceptance rate of 0.99 to reduce the likelihood of divergent transitions.

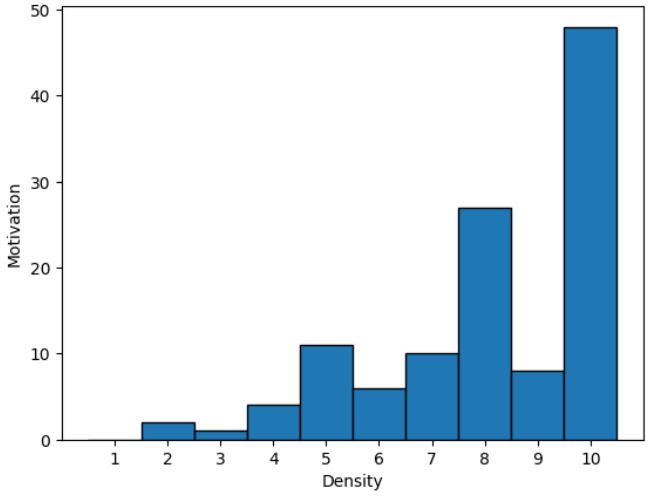


Figure 1. This histogram depicts the frequency with the students chose every level of motivation during the survey

Once the functional dependency between input and output variables was established, the function g was used to calculate the probability that students reach each motivation level. To accomplish this, we adopted the Monte Carlo method to simulate a broader input space than that available in the dataset [19]. This approach allows us to explore combinations of influencing factors not observed in the collected data, thereby providing a more comprehensive estimate of the risk of demotivation.

The probability that students achieve motivation level k for learning scientific computing is defined as follows:

$$P(y_i = k) \approx P(g(x_i) = k) = \int_{\mathcal{X}} P(g(x_i) = k \,|\, x_i) P(x_i)\, dx_i, \tag{3}$$

where $P(x_i)$ is the probability density function of the input variables.

Assuming that each component of x_i is uniformly distributed, i.e., $x_{ij} \sim \mathcal{U}(0, 1)$ for $j = 1, \ldots, D$, the probability density function $P(x_i)$ is uniform. Therefore, Equation (3) is rewritten as:

$$P(y_i = k) \approx P(g(x_i) = k) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(g(x_i) = k), \tag{4}$$

where N is the number of vectors x_i, whose components are random numbers uniformly distributed. Moreover, $\mathbf{1}(u) = 1$ if u is true, and $\mathbf{1}(u) = 0$ otherwise.

The value of N is chosen based on the standard error (SE), which is calculated as:

$$SE = \frac{\sigma}{\sqrt{N}}, \tag{5}$$

where σ is the standard deviation of the calculated probabilities. The value of N is increased iteratively until the SE decreases to a tolerable threshold.

The *Absolute Risk (AR)* of demotivation given the factor jth is not guaranteed is defined as follows:

$$AR(y_i < 5 \mid x_{ij} < 0.5) = \int_{\mathcal{X}} \frac{P(y_i < 5, x_{ij} < 0.5)}{P(x_{ij} < 0.5)}\, dx \tag{6}$$

Similarly, the AR of demotivation given the factor jth is guaranteed is defined as follows:

$$AR(y_i < 5 \mid x_j \geq 0.5) = \int_{\mathcal{X}} \frac{P(y_i < 5, x_{ij} \geq 0.5)}{P(x_{ij} \geq 0.5)}\, dx \tag{7}$$

The *Relative Risk (RR)* is defined as the ratio of these two quantities:

$$RR(y_i < 5 \mid x_{ij} < 0.5) = \frac{AR(y_i < 5 \mid x_{ij} < 0.5)}{AR(y < 5 \mid x_j \geq 0.5)} \tag{8}$$

Using the Monte Carlo method the $AR(y_i < 5 \mid x_{ij} < 0.5)$ is calculated as follows:

TABLE I. INPUT VARIABLES ASSOCIATED TO THE FACTORS THAT INFLUENCE THE STUDENT'S MOTIVATION IN SCIENTIFIC COMPUTING COURSES

Input Variable	*F-statistic*	*p-value*
The student's average grade in previous mathematics courses	0.43	5.16×10^{-1}
The extent to which the student has felt good about the course† $x_{i,1}$	26.17	1.27×10^{-6}
The extent to which the student has felt good about previous mathematics courses† $x_{i,2}$	24.68	2.38×10^{-6}
The extent to which the student has enjoyed the course† $x_{i,3}$	37.08	1.54×10^{-8}
The extent to which the student considers it imperative to study the course	1.08	3.02×10^{-1}
The extent to which the student considers it imperative to study mathematics courses	0.99	3.22×10^{-1}
The extent to which the student considers it wrong not to study the course	0.40	5.26×10^{-1}
The extent to which the student considers it wrong not to study mathematics courses	1.30	2.56×10^{-1}
The extent to which the student would like to recommend the course to other peers† $x_{i,4}$	37.27	1.43×10^{-8}
The extent to which the student perceives that the university provides them with up-to-date equipment† $x_{i,5}$	8.43	4.42×10^{-3}
The extent to which the course has encouraged students to study with classmates† $x_{i,6}$	29.49	3.17×10^{-7}
The extent to which the student has been encouraged to help classmates† $x_{i,7}$	29.09	3.74×10^{-7}
The extent of the student's current engagement in participating in course lessons† $x_{i,8}$	20.43	1.51×10^{-5}
The extent of the student's current engagement in attending course lessons	2.04	1.56×10^{-1}
The extent of the student's current engagement in making an additional effort to understand the course† $x_{i,9}$	27.31	7.82×10^{-7}
The extent of the student's current focus and engagement during course lessons† $x_{i,10}$	27.59	6.96×10^{-7}
The extent to which the student has been encouraged to study the course independently† $x_{i,11}$	31.37	1.48×10^{-7}
The extent to which the student has believed the course is useful for their professional life† $x_{i,12}$	20.64	1.37×10^{-5}
The extent to which the student has considered mathematics courses useful for their professional life† $x_{i,13}$	12.94	4.75×10^{-4}
The extent to which the student has believed that they possess the ability to learn mathematics† $x_{i,14}$	3.30	7.17×10^{-2}
The extent to which the student has believed that they have the ability to solve mathematics-related problems	0.93	3.37×10^{-1}
The extent to which the student has enjoyed to solve challenging mathematics-related problems similar to those addressed in the course	15.02	1.77×10^{-4}
The extent to which the student feels their secondary school preparation is insufficient for succeeding in mathematics courses	3.67	5.79×10^{-2}
The extent to which the student believes people have innate abilities for mathematics	1.96	1.64×10^{-1}
The extent to which the student believes learning success depends on the lecturer	3.80	5.36×10^{-2}
The extent to which the student believes learning success depends on the student	1.62	2.06×10^{-1}
The extent to which the student believes hard work is key to succeeding in the course† $x_{i,15}$	4.29	4.07×10^{-2}

†The input variable is selected for regression

TABLE II. MOTIVATION LEVELS OF THE STUDENTS WHO ANSWERED THE SURVEY

Motivation Level	*Number of Students*	*Proportion of the Sample*
2	2	1.71%
3	1	0.85%
4	4	3.42%
5	11	9.40%
6	6	5.13%
7	10	8.55%
8	27	23.08%
9	8	6.84%
10	48	41.02%
Total	117	100.00%

$$AR(y_i < 5 \mid x_{ij} < 0.5) \approx \frac{\sum_{i=1}^{N} \mathbf{1}(y_i < 5 \wedge x_{ij} < 0.5)}{\sum_{i=1}^{N} \mathbf{1}(x_{ij} < 0.5)} \quad (9)$$

Similarly, the $AR(y_i < 5 \mid x_{ij} \geq 0.5)$ is calculated as follows:

$$AR(y_i < 5 \mid x_{ij} \geq 0.5) \approx \frac{\sum_{i=1}^{N} \mathbf{1}(y_i < 5 \wedge x_{ij} \geq 0.5)}{\sum_{i=1}^{N} \mathbf{1}(x_{ij} \geq 0.5)} \quad (10)$$

Furthermore, the relative risk is calculated as follows:

$$RR(y_i < 5 \mid x_{ij} < 0.5) \approx \frac{\frac{\sum_{i=1}^{N} \mathbf{1}(y_i < 5 \wedge x_{ij} \geq 0.5)}{\sum_{i=1}^{N} \mathbf{1}(x_{ij} \geq 0.5)}}{\frac{\sum_{i=1}^{N} \mathbf{1}(y_i < 5 \wedge x_{ij} \geq 0.5)}{\sum_{i=1}^{N} \mathbf{1}(x_{ij} \geq 0.5)}} \quad (11)$$

Finally, if a factor exerts a negative influence on motivation (i.e., its associated weight is negative), the interpretation of "high" versus "low" values of that factor is reversed. To account for this, we compute the relative risk as $RR(y_i < 5 \mid x_{ij} \geq 0.5)$ rather than $RR(y_i < 5 \mid x_{ij} < 0.5)$. This adjustment ensures that the calculation consistently reflects the condition under which the factor increases the probability of demotivation.

IV. RESULTS

We estimated the weights of the function $g(x_i) = \beta^T x_i + \beta_0$ adopting the above-mentioned Bayesian regression model. The estimated values of the weights are reported in Table III. The largest weights correspond to students' satisfaction ($x_{i,4}$) and enjoyment ($x_{i,3}$) with the scientific computing course.

The negative weight for $x_{i,13}$ suggests that students who perceive mathematics courses as useful tend to be slightly less motivated in scientific computing courses. One possible explanation is that these students are primarily motivated by achieving high grades while considering mathematical knowledge mainly as a graduation requirement. Alternatively, they might feel demotivated because they prefer solving problems through analytical approaches (more common in

classical mathematics courses) in lieu of numerical methods, or heuristic, which are more common in scientific computing.

TABLE III. WEIGHTS OF THE PREDICTION FUNCTION ESTIMATED THROUGH THE BAYESIAN REGRESSION MODEL

Expected Weight	*97% CI*
$E[\beta_0] = 0.19$	[-1.224, 1.627]
$E[\beta_1] = 0.82$	[-0.718, 2.36]
$E[\beta_2] = 1.01$	[-0.38, 2.39]
$E[\beta_3] = 1.12$	[-0.44, 2.64]
$E[\beta_4] = 1.15$	[-0.45, 2.73]%
$E[\beta_5] = 0.45$	[-0.65, 1.55]
$E[\beta_6] = 0.52$	[-0.94, 1.94]
$E[\beta_7] = 0.89$	[-0.51, 2.33]
$E[\beta_8] = 0.56$	[-0.72, 1.84]
$E[\beta_9] = 0.83$	[-0.64, 2.31]
$E[\beta_{10}] = 0.79$	[-0.68, 2.25]
$E[\beta_{11}] = 1.01$	[-0.38, 2.38]
$E[\beta_{12}] = 0.35$	[-1.13, 1.817]
$E[\beta_{13}] = -0.38$	[-1.82, 1.08]
$E[\beta_{14}] = 0.27$	[-0.96 1.497]
$E[\beta_{15}] = 0.18$	[-1.148 1.499]

The Bayesian regression model adopted in this study achieved slightly better predictive performance than that reported by Caicedo-Castro et al. [1]. Our model attained a coefficient of determination (R^2) of 0.38 and a root-mean-squared error (RMSE) of 1.61, whereas the model in Caicedo-Castro et al. achieved an R^2 of 0.37 and an RMSE of 1.62. Although the improvement is marginal, it suggests that the Bayesian approach provides at least comparable, and potentially more robust, predictive accuracy.

Using the aforementioned function g, the results obtained from the Monte Carlo simulation revealed the most probable level is 4.98 with a standard error of 7×10^{-4}. Figure 2 illustrates how the simulation converges to this value as N increases. The estimate was obtained with 95% confidence (α = 0.05), yielding an interval [4.97, 4.98]. Since this value does not correspond to an actual motivation level, we rounded the result to the nearest even integer in halfway cases to ensure consistency with the discrete nature of the motivation scale (levels 1–10). The resulting probabilities of motivation levels are presented in Table IV.

TABLE IV. PROBABILITY OF EVERY MOTIVATION LEVEL CALCULATED WITH THE MONTE CARLO METHOD

Level	*Probability*
1	P(y = 1.0) = 1.22×10^{-4}%
2	P(y = 2.0) = 0.13%
3	P(y = 3.0) = 3.86%
4	P(y = 4.0) = 24.72%
5	P(y = 5.0) = 44.09%
6	P(y = 6.0) = 23.64%
7	P(y = 7.0) =3.46%
8	P(y = 8.0) = 9.78×10^{-2}%
9	P(y = 9.0) = 1.22×10^{-4}%

It is noteworthy that both the highest and lowest motivation levels have the smallest probabilities when the input variables are uniformly distributed across a broader space, as shown in Figure 3 and reported in Table IV. This suggests that extreme levels of motivation are less likely to occur under general conditions, and may instead arise from particular combinations of factors that are not equally represented in a uniform distribution. By contrast, the dataset indicates that students are predominantly highly motivated (see Table II in Section III), likely reflecting characteristics specific to the surveyed population rather than the broader input space.

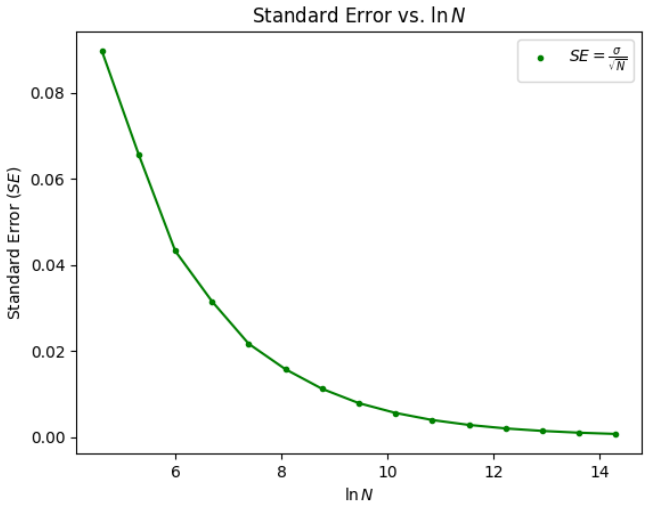


Figure 2. This chart shows how the standard error drops as the variable N is increased in the Monte Carlo simulation applied on the two-dimensional input space.

Satisfaction and enjoyment emerge as critical factors influencing the risk of demotivation, according to our simulation results. When the input space is explored through Monte Carlo methods, students who are unsatisfied with the scientific computing course and would not recommend it (i.e., input variable $x_{i,4}$) are more than twice as likely to become demotivated compared to satisfied students, with a relative risk of 2.41. In probabilistic terms, the simulation indicates that dissatisfaction raises the absolute risk of demotivation to 40.55%, while satisfaction reduces it to 16.85%. This corresponds to a risk difference of 23.7%, with a 95% confidence interval of [23.571, 23.838]. It is important to note that these figures do not reflect individual survey responses but instead arise from simulated projections across a broader range of possible student profiles. The results highlight how dissatisfaction can sharply elevate the probability of demotivation, underscoring the need to design course experiences that foster engagement and positive perceptions.

Moreover, the simulation results reveal that satisfaction and enjoyment, while related, are distinct factors influencing demotivation. Specifically, students who do not enjoy the scientific computing course (i.e., input variable $x_{i,3}$) are more than twice as likely to become demotivated compared to those who might be enjoying it, with a relative risk of 2.31. In absolute terms, lack of enjoyment increases the simulated risk of demotivation to 40.10%, while enjoyment reduces it to 17.31%. This yields a risk difference of 22.79%, with a 95% confidence interval of [22.653, 22.921]. Together with

the results for satisfaction, these findings underscore that both enjoyment and satisfaction independently contribute to mitigating demotivation, and that guaranteeing an enjoyable learning experience is as crucial as ensuring a satisfactory one.

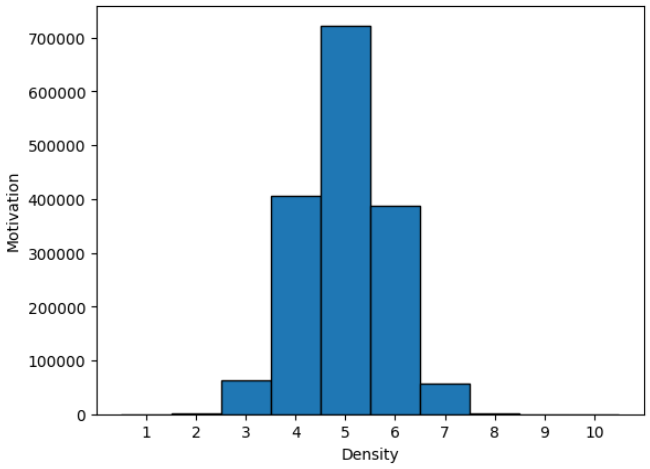


Figure 3. Histogram yielded through the Monte Carlo method. This shows the frequency with which the function g calculates each motivation level based on the random input variables.

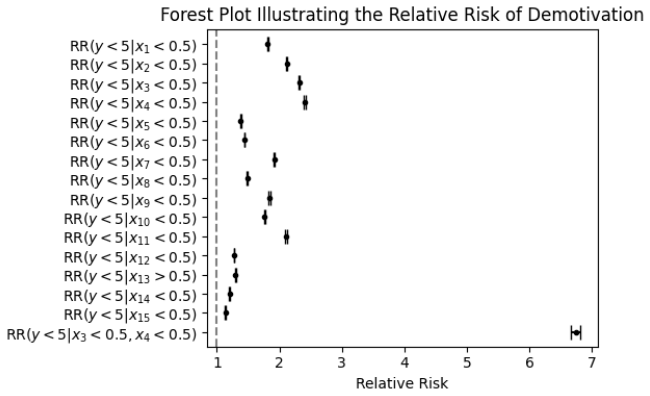


Figure 4. Forest plot showing the relative risk (RR) of failing the Physics II course. In all cases, the Wald test p-value is less than 0.05.

When both satisfaction and enjoyment are absent, the simulation shows a synergistic effect on the risk of demotivation. In this scenario, the absolute risk rises sharply to 54.54%, compared to only 8.09% among students who are both satisfied and enjoy the course. This corresponds to a risk difference of 46.45%, with a 95% confidence interval of [46.272, 46.619]. Put differently, the simulation indicates that students lacking both satisfaction and enjoyment are more than eight times as likely to become demotivated as their peers who experience both.

This combined effect is considerably stronger than the impact of each factor in isolation: dissatisfaction alone increases the absolute risk to 40.55%, while lack of enjoyment increases it to 40.10%. Thus, the simulation suggests that dissatisfaction and lack of enjoyment do not merely add up their effects, but instead interact to amplify the overall risk, underscoring the importance of addressing both factors simultaneously in the design of scientific computing courses.

Finally, the absolute and relative risks associated with the remaining factors are summarized in the forest plot shown in Figure 4 and detailed in Table V. This visualization provides a comparative view of how each factor contributes to the simulated risk of demotivation, allowing the relative importance of different influences to be assessed at a glance.

V. Discussion

The Monte Carlo method enables the estimation of demotivation risk in scientific computing courses without relying on direct experimentation with students, an approach that would be both unethical and impractical. This simulation-based framework therefore provides a valuable alternative for examining how different factors influence student motivation and for evaluating the prospective impact of educational policies. Whereas previous studies have primarily focused on predicting academic performance or identifying challenges associated with mathematical prerequisites, the present work extends this line of research by explicitly quantifying the risk of demotivation.

Fostering satisfaction, interest, and enjoyment in scientific computing requires designing courses in which students feel capable, perceive a clear sense of purpose, and experience continuous progress. Student satisfaction is enhanced when learning expectations are well defined and academic success is attainable; in this regard, transparent rubrics can help build confidence while reducing uncertainty and anxiety. In parallel, enjoyment increases when the learning experience is active, engaging, and meaningful, reinforcing students' intrinsic motivation and sustained involvement with course material.

Building on this contribution, the simulation results indicate that lecturers should prioritize fostering both satisfaction and enjoyment in scientific computing courses, as these factors consistently emerged as the strongest predictors of reduced demotivation. In other words, cultivating positive learning experiences may be as important as addressing cognitive challenges when designing effective courses.

Beyond satisfaction and enjoyment, the simulation results also indicate that encouraging independent study is a key factor in reducing the risk of demotivation. This finding suggests that lecturers should design learning environments that promote self-regulation, for example, by providing structured learning guides, formative assessments, and opportunities for collaborative problem-solving that still require individual accountability. At the same time, targeted interventions in prerequisite mathematics courses are needed to improve students' learning experiences. Redesigning these courses to emphasize conceptual understanding, applied problem-solving, and clear connections to scientific computing may help mitigate the negative impact of unfavorable prior experiences on student motivation.

From a pedagogical perspective, these findings underscore the importance of fostering self-regulated learning and im-

TABLE V. ABSOLUTE AND RELATIVE RISK OF DEMOTIVATION FOR LEARNING SCIENTIFIC COMPUTING.

Factor	Absolute Risk (%) exposed	Absolute Risk (%) unexposed	Risk Difference (%)	Relative Risk (%)	95% CI (Relative Risk)	95% CI (Risk Differences)
The student has not felt good about the course ($x_{i,1}$)	37	20.41	16.59	1.81	[1.803, 1.822]	[16.449, 16.721]
The student has not felt good about previous mathematics courses $x_{i,2<0.5}$	39	18.43	20.57	2.12	[2.105, 2.127]	[20.438, 20.708]
The student has not enjoyed the course ($x_{i,3}$)	40.10	17.31	22.79	2.32	[2.304, 2.329]	[22.653, 22.921]
The student would not recommend the course to other peers ($x_{i,4}$)	40.55	16.85	23.7	2.41	[2.384, 2.420]	[23.571, 23.838]
The student perceives the university lacks up-to-date equipment ($x_{i,5}$)	33.29	24.13	9.16	1.38	[1.373, 1.386]	[9.023, 9.298]
The course has not encouraged students to study with classmates ($x_{i,6}$)	33.9	23.53	10.37	1.44	[1.434, 1.448]	[10.237, 10.512]
The student has not been encouraged to help classmates ($x_{i,7}$)	37.78	19.65	18.13	1.92	[1.912, 1.932]	[17.990, 18.261]
The student lacks engagement to participate in course lessons ($x_{i,8}$)	34.35	23.07	11.29	1.49	[1.482, 1.497]	[11.150, 11.425]
The student lacks engagement to make an additional effort to understand the course ($x_{i,9}$)	37.22	20.22	17	1.84	[1.831, 1.850]	[16.860, 17.132]
The student struggles focus and lacks engagement during course lessons ($x_{i,10}$)	36.62	20.78	15.83	1.76	[1.753, 1.771]	[15.698, 15.971]
The student has been discouraged to study the course independently ($x_{i,11}$)	38.97	18.48	20.49	2.11	[2.097, 2.120]	[20.351, 20.621]
The student has believed the course is not useful for their professional life ($x_{i,12}$)	32.18	25.23	6.95	1.28	[1.269, 1.281]	[6.807, 7.083]
The student has considered mathematics courses useless for their professional life ($x_{i,13}$)	32.41	25.01	7.4	1.3	[1.290, 1.302]	[7.266, 7.542]
The student has believed that they lack the ability to learn mathematics ($x_{i,14}$)	31.35	26.07	5.28	1.2	[1.197, 1.209]	[5.146, 5.423]
The student believes hard work is not key to succeeding in the course ($x_{i,15}$)	30.55	26.87	3.68	1.14	[1.131, 1.142]	[3.538, 3.815]
The student has not enjoyed the course and would not recommend it to other peers ($x_{i,3}$ and $x_{i,4}$)	54.54	8.09	46.45	6.74	[6.668, 6.812]	[46.272, 46.619]

proving the design of prerequisite mathematics instruction. The simulation results reveal that neglecting factors such as students' encouragement to study independently and their prior experiences in mathematics courses is associated with an increased risk of demotivation, highlighting the need for coordinated instructional strategies that address both foundational preparation and autonomous learning skills.

To promote independent study, instructional approaches such as flipped classrooms, problem-based learning, and scaffolded assignments can be particularly effective, as they combine learner autonomy with structured pedagogical support. These methods encourage students to take responsibility for their learning while ensuring they receive ongoing guidance and formative feedback.

Flipped classrooms, or flipped learning, reverse the traditional instructional model in which lecturers deliver content during class time and students complete practice activities at home. In this approach, students engage with instructional materials (such as, e.g., readings, videos, or interactive resources) prior to class, allowing face-to-face time to be devoted to problem-solving, discussion, project work, and direct interaction with the lecturer. As a result, students assume a more active role in the learning process, which promotes deeper engagement by transforming class sessions into interactive rather than passive experiences. Moreover, flipped learning supports self-paced study, enabling students to revisit materials as needed and thereby gain greater control over their learning process.

Flipped learning can be effectively combined with problem-based learning, where class time is dedicated to addressing real, open-ended problems rather than listening to traditional lectures. In this integrated approach, students are motivated to acquire the concepts, theoretical knowledge, and practical skills necessary to solve the problems posed in class. At the same time, they strengthen their ability to engage in independent study and develop critical thinking skills through active inquiry and collaborative problem-solving.

Scaffolded assignments can be used to complement both flipped learning and problem-based learning. These assignments are designed to provide structured, temporary support that helps students gradually build the skills and competencies required to work independently. The educational concept of scaffolding is inspired by the construction scaffold, which supports workers during the building process and is progressively removed as the structure becomes self-sustaining.

In order to improve the design of prerequisite mathematics instruction for scientific computing, coursework should focus on preparing students to model, simulate, and reason algorithmically. Mathematics courses can be oriented around applied use cases relevant to scientific computing, such as linear algebra for analyzing the stability of numerical solvers, differential calculus for numerical optimization and differentiation, and integral calculus for Monte Carlo estimation. In this approach, mathematics is taught through algorithms rather than solely through symbolic manipulation, enabling students to implement methods, explore numerical error and conditioning, visualize results, and work with approximations, thereby strengthening the practical connections between mathematical theory and computational application.

Strengthening prerequisite mathematics courses requires conceptual teaching approaches that deliberately connect abstract ideas with practical applications in scientific computing. Instructional strategies such as contextualized problem-solving, interdisciplinary projects, and active learning techniques (e.g., peer instruction and inquiry-based exercises) can make mathematics more meaningful and directly relevant to students' future coursework. Together, these pedagogical approaches not only help reduce the risk of demotivation but also foster continuity between foundational mathematical training and applied scientific computing.

In contrast, students' beliefs show a relatively weak association with the risk of demotivation. Perceptions such as viewing hard work as unimportant for success in scientific computing, doubting one's ability to learn mathematics, or considering mathematics or scientific computing to have little practical value are associated with comparatively low relative risks. Moreover, the corresponding differences in absolute risk are substantially smaller than those observed for other factors, indicating a limited practical impact of these beliefs on demotivation.

The weak association observed between students' beliefs and demotivation suggests that these beliefs may function as mediating factors rather than direct determinants of motivational outcomes. Negative beliefs concerning effort, ability, or the value of mathematics and scientific computing may develop in response to prior learning experiences, including academic difficulty, ineffective instructional practices, or early failure. Consequently, experiential conditions may shape learning outcomes first, followed by adjustments in students' beliefs, with demotivation emerging thereafter. Because the regression model includes variables with stronger effects (such as, e.g., satisfaction and enjoyment) the independent contribution of beliefs is attenuated, resulting in comparatively low relative risks and small absolute risk differences associated with these factors.

Limited access to up-to-date equipment also exhibits a relatively low relative risk and a small absolute risk difference. In contemporary educational contexts, this factor is not a dominant barrier to student motivation, as the tasks typically performed in scientific computing courses do not require high-performance hardware. Consequently, inadequate equipment alone is unlikely to constitute a substantial contributor to demotivation.

When most students have adequate access to functional equipment, or when mitigation strategies are widely available through appropriate instructional design, the exposed group (those who truly lack usable hardware) becomes small and heterogeneous. This statistical compression reduces discriminative power and biases both relative and absolute risk estimates toward modest values. Furthermore, scientific computing courses frequently make use of platforms such as Google Colab, which allow students to access sufficient computational resources even from low-spec devices, including smartphones.

Finally, it is important to acknowledge the methodological

scope of this study. Expanding the dataset to include additional explanatory variables and a larger number of observations, as well as adopting more sophisticated models capable of capturing nonlinear relationships between inputs and outcomes, could further improve predictive performance. These enhancements may reduce the root-mean-square error and increase the coefficient of determination, thereby strengthening the robustness and validity of the simulation results. Overall, by integrating computational simulation with pedagogical analysis, this work contributes methodologically to educational research while offering practical guidance for enhancing student motivation in scientific computing courses.

VI. Conclusion and Perspectives

We employed the Monte Carlo statistical method to estimate the absolute and relative risk of student demotivation in undergraduate scientific computing courses within the Systems Engineering program at the University of Córdoba (Colombia). This approach enabled the simulation of a large number of scenarios that cannot be examined empirically, thereby allowing exploration of a broad range of values for the independent variables representing factors associated with student motivation, the dependent variable of the study. The relationships between these factors and motivation were quantified using Bayesian regression, which provided the functional dependencies used as inputs for the simulation.

The results of the Monte Carlo simulations indicate that the factors most strongly associated with increased risk of demotivation are: i) students' satisfaction with the scientific computing course, ii) the level of enjoyment perceived by students during the course, iii) students' experiences in prerequisite mathematics courses, and iv) discouragement toward independent learning.

A hybrid pedagogical approach that integrates flipped instruction, scaffolded assignments, problem-based teamwork, and frequent formative feedback may offer an effective course design for enhancing student motivation while maximizing enjoyment and satisfaction in scientific computing. Such an approach aims to reduce anxiety, support the development of competence, increase the perceived relevance of learning activities, foster autonomy, and strengthen social connection, thereby improving the likelihood that students remain motivated and engaged throughout the course.

Additional simulation findings reveal that students' beliefs about effort, personal ability, and the relevance of mathematics and scientific computing are associated with a relatively low risk of demotivation. Similarly, the availability of up-to-date equipment at the university shows only a minor association with demotivation risk, suggesting that these factors exert a limited practical influence compared with the primary predictors identified in the study.

For future research, we shall conduct intervention testing and instructional design evaluations through controlled trials that compare pedagogical approaches such as scaffolded instruction, flipped learning, and problem-based learning. These studies will aim to verify whether and to what extent these strategies effectively enhance student motivation in scientific computing courses.

So far, we have assumed a linear relationship between student motivation and its influencing factors. In future work, we plan to explore nonlinear regression models in which the independent variables are mapped into higher-dimensional feature spaces. Such approaches may better capture complex interactions among factors, potentially increasing the coefficient of determination while reducing the root-mean-squared error.

Additionally, we shall apply dimensionality reduction techniques (such as, e.g., principal component analysis, matrix factorization, and autoencoders) to identify latent factors underlying student motivation. These methods may provide more compact and informative representations of the independent variables by filtering noise and capturing the most relevant structure in the data, thereby improving the performance and interpretability of the regression models.

Another extension of this research is the development of Monte Carlo–based causal simulations to move beyond risk association toward explicit counterfactual policy evaluation. Under this approach, a probabilistic causal structure (derived from methods such as Bayesian causal networks, structural equation modeling, or quasi-experimental estimators) is specified and used as a generative model of the educational system. Large numbers of simulated student trajectories can then be sampled under alternative interventional scenarios (e.g., enhanced instructional quality or the introduction of scaffolded assignments) using the logic of do-calculus. This framework enables the estimation of causal quantities such as the expected reduction in demotivation, changes in satisfaction levels, and heterogeneity of intervention effects across student subgroups, all while accounting for uncertainty.

Monte Carlo causal simulation is particularly well suited to educational research because it can capture the nonlinear dynamics, mediation pathways, interaction effects, and threshold behaviors inherent to learning processes—phenomena that are difficult to isolate using closed-form regression models alone. This approach enables the evaluation of complex, multicomponent interventions rather than single-factor manipulations, for example by assessing whether moderate, coordinated improvements across several instructional dimensions yield larger motivational gains than major changes in only one area. In addition, sensitivity analyses can be directly integrated into the simulation framework to quantify how unmeasured confounding, measurement error, or parameter uncertainty propagate into causal predictions, thereby providing more transparent and realistic uncertainty bounds than conventional point estimates.

Finally, implementing this framework would enable a shift from descriptive modeling toward decision-support analytics for curricular and pedagogical design. Simulated policy experiments could identify which combinations of instructional interventions are most cost-effective, determine which student profiles benefit most from targeted support, and reveal points at which diminishing returns occur.

ACKNOWLEDGMENT

Caicedo-Castro thanks the Lord Jesus Christ for blessing this project. The authors thank Universidad de Córdoba in Colombia for supporting this study. They also thanks all students who collaborated by answering the survey conducted for collecting the dataset used in this study. Finally, the author thanks the anonymous reviewers for their comments that contributed to improve the quality of this article.

REFERENCES

[1] I. Caicedo-Castro, O. Vélez-Langs, and R. Castro-Púche, "Using the Monte Carlo Method to Estimate Student Motivation in Scientific Computing", in *PATTERNS 2025: The Seventeenth International Conferences on Pervasive Patterns and Applications*, ser. International Conferences on Pervasive Patterns and Applications, Valencia, Spain: IARIA: International Academy, Research, and Industry Association, 2025, pp. 15–22, ISBN: 978-1-68558-263-0.

[2] I. Caicedo-Castro, M. Macea-Anaya, and S. Rivera-Castaño, "Early Forecasting of At-Risk Students of Failing or Dropping Out of a Bachelor's Course Given Their Academic History - The Case Study of Numerical Methods", in *PATTERNS 2023: The Fifteenth International Conference on Pervasive Patterns and Applications*, ser. International Conferences on Pervasive Patterns and Applications, Nice, France: IARIA: International Academy, Research, and Industry Association, 2023, pp. 40–51, ISBN: 978-1-68558-049-0.

[3] I. Caicedo-Castro, "Course Prophet: A System for Predicting Course Failures with Machine Learning: A Numerical Methods Case Study", *Sustainability*, vol. 15, no. 18, 2023, 13950. DOI: 10.3390/su151813950.

[4] I. Caicedo-Castro, "Quantum Course Prophet: Quantum Machine Learning for Predicting Course Failures: A Case Study on Numerical Methods", in *Learning and Collaboration Technologies*, P. Zaphiris and A. Ioannou, Eds., Cham: Springer Nature Switzerland, 2024, pp. 220–240, ISBN: 978-3-031-61691-4. DOI: 10.1007/978-3-031-61691-4_15.

[5] I. Caicedo-Castro, "An Empirical Study of Machine Learning for Course Failure Prediction: A Case Study in Numerical Methods", *International Journal on Advances in Intelligent Systems*, vol. 17, no. 1 and 2, pp. 25–37, 2024.

[6] L. Ayebale, G. Habaasa, and S. Tweheyo, "Factors Affecting Students' Achievement in Mathematics in Secondary Schools in Developing countries: A Rapid Systematic Review", *Statistical Journal of the IAOS*, vol. 36, pp. 1–4, 2020. DOI: 10.3233/SJI-200713.

[7] M. Gómez-García, H. Hossein-Mohand, J. M. Trujillo-Torres, H. Hossein-Mohand, and I. Aznar-Díaz, "Technological Factors That Influence the Mathematics Performance of Secondary School Students", *Mathematics*, vol. 8, no. 11, 2020, 1935, ISSN: 2227-7390. DOI: 10.3390/math8111935.

[8] J.-M. Trujillo-Torres, H. Hossein-Mohand, M. Gómez-García, H. Hossein-Mohand, and F.-J. Hinojo-Lucena, "Estimating the Academic Performance of Secondary Education Mathematics Students: A Gain Lift Predictive Model", *Mathematics*, vol. 8, no. 12, 2020, 2101, ISSN: 2227-7390. DOI: 10.3390/math8122101.

[9] M. Maamin, S. M. Maat, and Z. H. Iksan, "The Influence of Student Engagement on Mathematical Achievement among Secondary School Students", *Mathematics*, vol. 10, no. 1, 2022, 41, ISSN: 2227-7390. DOI: 10.3390/math10010041.

[10] A. Brezavšček, J. Jerebic, G. Rus, and A. Žnidaršič, "Factors Influencing Mathematics Achievement of University Students of Social Sciences", *Mathematics*, vol. 8, no. 12, 2020, 2134, ISSN: 2227-7390. DOI: 10.3390/math8122134.

[11] E. Martinez-Villarraga, I. Lopez-Cobo, D. Becerra-Alonso, and F. Fernández-Navarro, "Characterizing Mathematics Learning in Colombian Higher Distance Education", *Mathematics*, vol. 9, no. 15, 2021, 1740, ISSN: 2227-7390. DOI: 10.3390/math9151740.

[12] J. Park, S. Kim, and B. Jang, "Analysis of Psychological Factors Influencing Mathematical Achievement and Machine Learning Classification", *Mathematics*, vol. 11, no. 15, 2023, 3380, ISSN: 2227-7390. DOI: 10.3390/math11153380.

[13] S. Batista-Toledo and D. Gavilan, "Student Experience, Satisfaction and Commitment in Blended Learning: A Structural Equation Modelling Approach", *Mathematics*, vol. 11, no. 3, 2023, 749, ISSN: 2227-7390. DOI: 10.3390/math11030749.

[14] M. Charalambides, R. Panaoura, E. Tsolaki, and S. Pericleous, "First Year Engineering Students' Difficulties with Math Courses- What Is the Starting Point for Academic Teachers?", *Education Sciences*, vol. 13, no. 8, 2023, 835, ISSN: 2227-7102. DOI: 10.3390/educsci13080835.

[15] T. T. Wijaya, B. Yu, F. Xu, Z. Yuan, and M. Mailizar, "Analysis of factors affecting academic performance of mathematics education doctoral students: A structural equation modeling approach", *International Journal of Environmental Research and Public Health*, vol. 20, no. 5, 2023, 4518, ISSN: 1660-4601. DOI: 10.3390/ijerph20054518.

[16] M. D. Hoffman and A. Gelman, "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo", *Journal of Machine Learning Research*, vol. 15, no. 47, pp. 1593–1623, 2014.

[17] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006, ISBN: 0387310738.

[18] O. Abril-Pla et al., "PyMC: a Modern, and Comprehensive Probabilistic Programming Framework in Python", *PeerJ Computer Science*, vol. 9, no. 1516, 2023, ISSN: 2376-5992. DOI: 10.7717/peerj-cs.1516.

[19] N. Metropolis and S. Ulam, "The Monte Carlo Method", *Journal of the American Statistical Association*, vol. 44, no. 247, pp. 335–341, 1949, ISSN: 01621459, 1537274X.

Culture, Agendas and the Effect of Social Media on Malaysian Politics - A Literature Review

Sayantan Bhattacharya*, Nitin Agarwal*†
*COSMOS Research Center, University of Arkansas at Little Rock, Arkansas, USA
†International Computer Science Institute, University of California, Berkeley, California, USA
e-mail: {sbhattachary, nxagarwal}@ualr.edu

***Abstract*—Malaysia is a multicultural nation with a complex political system. Several variables, including racial, religious, and economic diversity, have altered the political landscape. The political climate has been unstable in recent years, especially around general election time. The Malaysian government has been criticized for influencing the media to further its own goals. In some instances, bloggers and journalists who had spoken out against the administration were arrested to stifle free speech. In addition to possible Chinese intervention in internal politics and ethnic unrest that various political groups occasionally stoke to win support, there is also the government's censorship. This essay examines Malaysian studies from the perspective of social media analysis. How social media has reshaped Malaysia, several political scenarios, and the general election. Without a doubt, the internet and social media have become increasingly crucial in Malaysian political discourse. The total internet penetration in Malaysia increased from 1,718,500 in 2008 to 5,839,600 in 2012. Malaysia has 13 million Facebook users and two million Twitter users out of a population of 29 million. In Malaysia, 64 percent of Facebook users are aged between 18 and 34, while 62 percent of total unique Internet visitors are aged between 15 and 34. With approximately 30 percent of Malaysia's 13.3 million registered voters for the GE13 being first-time voters, the Internet played a critical role. Before GE13, social media had played a significant part in various mobilization actions as well as the proliferation of anti-establishment propaganda shared by young urbanites. Reuters Institute Digital News Report of 2022 estimates Malaysia's internet penetration to be around 89 percent. Social media sites were the second most popular news sources at 75 percent. Fifty-two percent of respondents shared the news through social media sites with Facebook (52 percent), WhatsApp (47 percent) and YouTube (39 percent) being the top three social media platforms for news. Malaysia's situation is complex, multifaceted, and fraught with challenges. In this survey paper, we have made an effort to discuss all these traits and difficulties from the perspective of social media and online behavioral studies. It will provide an outline of the research that has been done for these studies.**

***Keywords*-Social media; China Influence; Government censorship of free speech; Political Actors; Voting Patterns; Human Rights; Media Freedom; Political Polarization; Economic Development.**

I. Introduction

Malaysia's government does not exactly adhere to the continuum between authoritarianism and democracy. Many parties compete in Malaysia's frequent and regular elections, which are far from free and fair and effectively lead to one-party control. Malaysia has a history of controlling and regulating the media through legislative measures, with the government taking actions to suppress the free speech of citizens, including arresting independent journalists and bloggers. There has also been considerable effort to manipulate social media and divert public attention.

With its multi-ethnic and multi-religious composition, Malaysia and its politics have been widely studied in the literature, especially in its approach to democratic governance. Malaysia's politics is complex and shaped by its unique history, diverse cultures, and practices. The country is a federal constitutional monarchy with a parliamentary democracy system, with a monarch serving as head of state and a prime minister serving as the head of government.

In this paper, we will review existing literature to analyze different themes and their impact on Malaysian politics. The themes covered in this paper include voting patterns across several population subsets, the evolution of social media as a political tool, the actions, actors, and agendas of ruling regimes in Malaysian politics, and China's influence in Malaysia. Our goal is to provide a comprehensive, consolidated summary of pivotal moments in Malaysian politics, identify quantitative and qualitative gaps in existing research, and lay a foundation for future research on Malaysian politics.

While this survey focuses on Malaysian politics, it is situated within a broader body of work on how social media structures online communities and shapes their cohesion. Prior research on COVID-19 vaccine discourse on X has shown that toxicity can measurably alter community structure—spikes in toxic speech are associated with drops in modularity and clustering coefficients, indicating that communities become more fragmented over time. In the Malaysian context, our review and CFSA-based network analysis already reveal tightly knit focal structures and platform-specific clusters around parties, media, and activists, as well as episodes of ethnic and religious hate speech, particularly on TikTok. Future work can therefore extend this survey by explicitly modeling Malaysian political communities as dynamic networks and examining how the spread of toxic or hateful content affects their cohesion, polarization, and fragmentation over time, paralleling the toxicity–community dynamics framework used in the earlier study [1].

1) Structure of the Paper: From here on, the paper is organized into 4 main sections as follows: Section II explains how we selected the documents used in this study. In Section III, we will analyze data such as keywords, authors and trends about Malaysian politics and this paper. We will also summarize our findings and share possible future research directions.

II. Analysis of The Malaysian Political Landscape

A. Voting Patterns

This section will analyze subgroups within the Malaysian populace (gender, age, ethnicity, and urban versus rural dwellers) and identify common voting patterns among them.

B. Gender

The Department of Statistics Malaysia (DOSM) in 2022 estimated the Malaysian population to be at 32.7 million with 17 million being male and 15.7 million being female [2]. In its 2021 report, the DOSM stated that the gender equality rate was 71.4 percent as of 2020. Malaysia maintained its 8th ranking in East Asia and was ranked 74th in the world, based on its Malaysian Gender Gap Index score in 2020 [3].

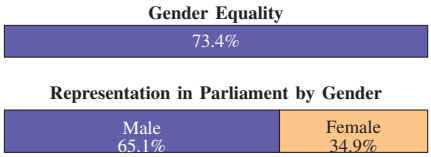


Figure 1: Gender Equality and Representation in Parliament by Gender (Data Source: World Bank (2020), Malaysia Department of Statistics (2020)

According to the World Bank, as of 2019, the adult literacy rate in Malaysia was lower among women at 93.6 percent compared to men at 96.2 percent. The labor force participation rate was also lower for women at 51.2 percent and 77.6% for men. The World Bank also estimates that women held 14.9 percent of parliamentary seats as at 2020 [4].

To understand the low participation rate of women in Malaysian politics, it is helpful to assess how the gender structure in Malaysia has changed over the years. The current role of women in Malaysia is a mix of traditional Malaysian customs, Islamic influences and recent socio-political events and while the government makes efforts to decrease gender inequality, the results of these efforts are minimal especially when it comes to women holding power in politics and the economy as a whole. (Kennedy 2002). One of such government efforts was the National Policy for Women of 1989 which had objectives that sought to ensure equal access and integration of women in all sectors of national development [5]. Another more recent effort by the Malaysian government in increasing women participation in politics is its endorsement of the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW), the United Nations treaty on women's equality [6].

The Malay woman as represented in literature has evolved over the years; from the historical Malay woman who had relative autonomy [7], idealized in female rulers like Che Siti Wan Kembang and Ratu Shafiuddin [8], which evolved to the factory girl who had lost some autonomy as a result of capitalism and patriarchy, who further developed to the modern Malay-Muslim woman, embodied by women such as Wan Azizah; veiled and devout, educated, loyal to family with a degree of independence [7].

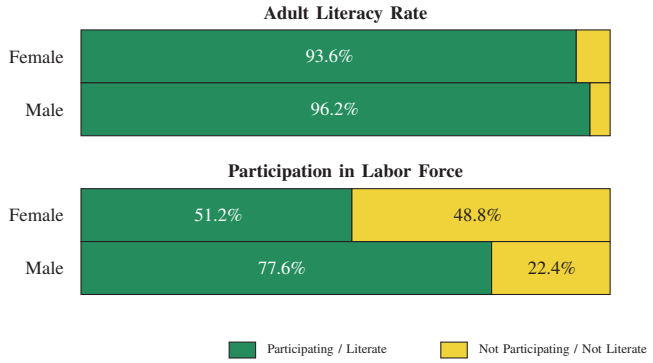


Figure 2: Adult literacy and labor force participation by gender.

Malaysian society struggles to juxtapose its traditional views about women with the increasing outcry for gender equality and women's acceptance into traditionally male-dominated spaces. Although women now have careers and support their families economically, they are still expected to prioritize domestic responsibilities. This expectation is often reinforced through concepts such as taat (obedience) and symbolism, which are frequently deployed to remind women of their place and roles in society [9].

In a 2009 study conducted by Mellstrom, female students expect to balance domestic responsibilities with career responsibilities, choosing careers that allow them to do so easily, with some anticipating early retirement to focus on their families [10] fully. A study by Noor (1999) highlights the cultural differences in how Chinese and Malay women approached their careers. Confucianism prioritizes family welfare over individual welfare, and this impacts Chinese women's attitudes towards work in that they think of work as an economic necessity needed to contribute to the family. On the other hand, Malay women expressed that while work was for family gains, material gains are not the most essential thing in life [11].

As female education became more prevalent and Malaysian women were exposed to ideas about gender equality, they became increasingly conscious of their marginalization. They had to confront hegemonic processes in their desire to participate more in society [12]. For example, Malay-Muslim women, in attempts to navigate the tension between Islamic beliefs and the expectations of the modern world, have adopted various tactics, in ways similar to tactics described by Kandiyoti (1988) as patriarchal bargaining. Some have adopted the term 'womanist' rather than 'feminist' to avoid the negative cultural implications associated with the 'feminist' term [7]. At the same time, some have used the veil to garner public acceptance as they make career moves in politics and public management [13]. It is against this cultural backdrop that Malaysian women participate in politics.

Malaysian politics is divided along ethnic and religious lines, and politicians must identify themselves with their

ethnic group and/or religion. Ethnicity-based politics takes precedence over gender politics, as reflected in the division of women's involvement in politics along ethnic party lines [10]; consequently, their participation in politics is viewed through the lens of ethnicity rather than gender [14].

In terms of political preferences and voting, a survey of Malaysian women voters after the 14th General Election found that women made voting decisions based on multiple factors, including party identification, their demographics, and the information they had gathered about the candidates. Age was a significant factor as older women tended to vote for key leaders while younger female voters prioritized issues over candidates or parties. The voter's ethnic group also mattered, as Indian women were shown to prefer issue-based politics while Malay and Chinese women favored candidate/party-based politics. Malay women also showed reduced support for increased female participation in politics compared to women from other ethnic groups [15].

Any benefits or support that women politicians receive are usually based on their ethnicity rather than their gender. For example, Malay women in politics receive preferential treatment based on their Malay identity and the associated benefits accorded to Malays, rather than on their merits as women politicians [13]. This ethnicity-based view of women in politics has led to the relegation of women's agenda to the background as a focus on women's agenda could be penalized by public condemnation [14], therefore preventing the development of a collective affinity between women.

Women's participation in politics is often relegated to helper roles such as mobilizing other women to support the party's candidates and taking care of logistics [16]. They have also had to downplay their roles to continue receiving public support. For example, in 2001, the woman minister of the Ministry of Women and Family Development had to emphasize that the ministry, previously known as the Ministry of Women's Affairs, would prioritize advocating for women's traditional roles to prevent upsetting the electorate. To receive nominations and support from their parties or electorates, women often adopt masculine traits or nicknames, risk being perceived as selfish for prioritizing politics over family, and navigate past traditionally male party gatekeepers [17].

The story is not too different in civil society movements. While women have been able to assert their voices through civil society movements [6], they still struggle to bridge ethnic and gender divides with respondents to a survey by Ng (2010) sharing difficulties they have faced such as derision on their pushing women's agenda or criticisms on their choice of dressing [18].

Women issues have also been co-opted by political parties and coalitions to gain points in electoral battles. When Wan Azizah Wan Ismail, Anwar Ibrahim's wife, became prominent following her husband's travails and the creation of the Reformasi, Parti Islam Se-Malaysia (PAS), that did not field any woman candidate for elections and whose leaders had previously dissuaded women from public roles, rallied behind her in a show of solidarity. Barisan Nasional (BN) also championed women's rights to distract the public from its history of corruption and poor governance.

In recent times, the Pakatan Harapan Coalition (PH) had, as part of its GE14 campaign, promised to put 30 percent of women in parliament. Shortly after elections, a group of women activists called for the coalition to uphold its promises, a call that culminated in a Twitter social media campaign known as the 30 peratus (30 percent) campaign. The campaign continued for months despite counter-calls asking women to earn their rights rather than being given quotas, and that the campaigners should wait for the 'right time' (Yoong 2019), further underscoring the prevailing view of Malaysian society towards women.

Even as Malaysia seems to be having more success towards more women participation in politics, as seen in the GE14 with the election of the highest percentages of female lawmakers among other women victories, there are still issues around low female candidates nominations, a situation identified by Yeong (2018) as a barrier for women party members, midst socio-cultural other problems.

C. Ethnicity

The Department of Statistics Malaysia [2] estimates Malaysia's population to be around 32.7 million people. The major ethnic groups comprising this population are the Malays and the Indigenous people (Orang Asli, Dayak, Anak Negeri etc.) collectively known as the Bumiputera, the Chinese, the Indians, and other ethnic groups.

To fully understand how ethnic dynamics influence Malaysian politics, it is essential to consider the legacy of British colonization in Malaysia. Britain colonized Malaysia until its independence in 1957, and to date, the effects of this colonization are still visible in Malaysia. The British employed the indirect rule style to manage their affairs in Malaysia, meaning that traditional Malay leaders retained their figurehead status but had limited political authority, despite participating in state administration. Colonization also separated the rural peasants from the local elites, reserving modern education and economic activities for the elites and urban areas [19].

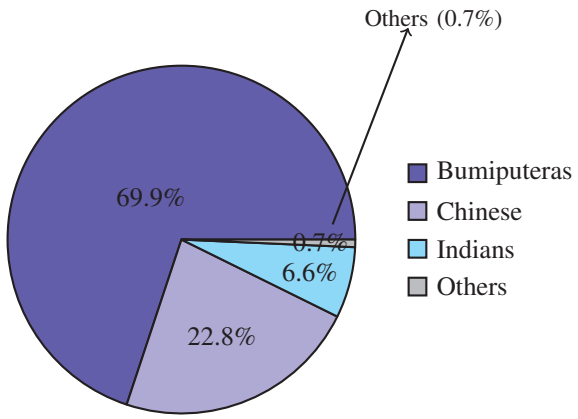


Figure 3: Ethnic Composition in Malaysia

To support the increased economic activities in Malaysia, the British facilitated labor immigration from China and

India, a situation that altered the demographics of Malaysia, especially the western half of the peninsula, which is regarded as the Federated States [20]. Due to this influx of immigrants, three distinct labor lines started to emerge where the Malays continued their traditional economic activities of farming, the Indians worked in the rubber estates in rural settlements, and the Chinese worked in mines and traded in the wealthier urban areas [21].

The British also accorded some form of status protection to the Malays, giving them preference for civil/public roles in the military, civil service, and police force, and also protecting their occupation of the land [22]. These jobs paid less and were unattractive to the other groups [23], but were politically important, reinforcing the political supremacy of the Malays. Land protection also created distinct living areas for the ethnic groups, with the Malays living in the rural areas, and the Chinese and Indians (except Indian plantation workers) living in the urban areas. The economic advantage the Chinese gained from their exposure to more lucrative jobs, the special preference for Bumiputeras, and the high contrast between the three cultures became primary sources of tension in Malaysia. This tension culminated in several incidents, notably the Sino-Malay riots in May 1969.

The British exercised less control over the unfederated states of Malaysia, particularly the northern states of Kelantan, Terengganu, Kedah, and Perlis, which excluded these areas from experiencing the same economic and demographic changes as the other areas, thereby allowing them to retain their distinct features [20].

D. Ethnic Groups and Voting Patterns in Malaysia

1) Malays: The Malays are traditionally Muslim, rural inhabitants who value brotherhood and mutual assistance [24]. They prioritize cooperation for the common good, embracing the solidarity and sense of belonging as taught by Islam [25]. By independence, most Malays were still dominantly in the rural areas working as farmers or fishers, and occupying civil and government roles. Their political attitude is influenced by feudalism, traditional western values, and Islam, culminating in a deference and uncritical acceptance of authority by the citizens and expectations of respect and electoral support by the political leaders [23].

When planning their exit and deciding the future of the Malayan Union (now Malaysia), the British attempted to create a more democratic government, but met resistance from the Malays, who believed that such an arrangement would downgrade their special position [22]. The special consideration afforded to the Malays has influenced much of Malaysia's politics. It influenced the formation of the United Malays National Organisation (UMNO) in 1946, which aimed to advocate for Malay interests and was mainly composed of Malay aristocrats [22].

The 1957 Constitution of Malaysia further reinforced Malay special privileges, retaining Malay as the official language, with Article 153 of the Constitution specifying a reserved proportion of public service positions for Malays, among other special benefits.

UMNO went on to be a founding member of the Barisan Nasional (BN) coalition, and the support they received from Malays, based on their advocacy for Malay interests, informed the political dominance of Barisan Nasional for the decades after independence.

2) Chinese: The Chinese are considered the most heterogeneous of the ethnic groups in Malaysia in terms of language and religion, and this has influenced their political attitude. They are the wealthiest ethnic group in Malaysia, aided by their presence in the wholesale/retail (27.7 percent), manufacturing (26.4 percent), and other services (13.8 percent) sectors [26]. They are concentrated in the more developed urban areas with access to better educational opportunities, which has increased their level of political knowledge. The Chinese have Confucian values that encourage the vocal airing of grievances and the seeking of redress when wronged, which means they were vocal against perceived discrimination in Malaysia. This situation has created an impasse for the Chinese party, Malaysian Chinese Association(MCA), founded in 1949, whereby they feel obligated to advocate against Chinese discrimination, but risk alienating the Malay electorate when they do so, but also when they don't advocate for the Chinese, they alienate the Chinese community and are decried as sellouts [24].

This lack of a communal voice in MCA may explain the divided loyalties of the Chinese, most of whom support MCA, with some supporting opposition parties [23]. MCA went on to join Barisan Nasional and the Chinese support they received also contributed to BN's dominance in politics.

Indians: The Indians are the smallest of the three major ethnic groups, settled in urban, rural, and suburban areas. Composed mainly of Tamil-speaking Hindus, they did not have a majority in any Malaysian constituency [27].

Culturally, Indians are expected to yield to family and caste authority and are more concerned with their moral development than with seeking redress for injustice. This, alongside their smaller numbers in Malaysia, may explain why scholars generally see the Indian ethnic group as passive and amenable to authorities.

Without the special treatment received by Malays, the economic advantage of the Chinese, and with the decline of the Malaysian agricultural sector, the Indians are the poorest of the three major ethnic groups [28]. The combination of these factors meant that the Indians do not have significant political bargaining power in Malaysian politics, further marginalizing them.

A few Indians support the opposition but most support the Malaysian Indian Congress (MIC) which was part of the BN coalition, allowing them to access the benefits associated with belonging to the ruling regime [22].

The continued ethnic, religious and economic marginalization culminated in the creation of HINDRAF (Hindu Rights Action Force) in 2006 to advocate for Indian interests, which mobilized the HINDRAF rally of 2007 [27].

3) The 1Malaysia Concept: Issues such as the lack of cultural and regional integration and the economic and political imbalance among all the ethnic groups have prevented

genuine pluralism from occurring in Malaysia, as racial tensions and imbalances continue to persist. To address this, Malaysia, under Prime Minister Razak, introduced the 1Malaysia concept to create a shared identity and nationalistic spirit among Malaysians of all ethnic groups [29]. However, researchers have found that the concept is not fully embraced by Malaysians. According to Noor et al. (2013), the Malays see it as a way to undermine their special rights, while the non-Malays see it as a political strategy [30] and participants in a study by Harris et al. (2020) described the concept as an 'advertising slogan' with little change being effected on inter ethnic relations [31].

The Malaysian government continues to make efforts towards a more integrated Malaysia; however, these efforts seem to be yielding little results. A 2022 report by PUSAT KOMAS about racism in Malaysia found an increase in incidents of racial discrimination and racism in Malaysia, citing 82 incidents, a record high over the previous record of 76 incidents in 2018 and higher than the 55 incidents recorded in 2021. Most of these incidents were related to politics and may be due to 2022 being an election year in Malaysia [32]. This shows that racial politics continue to persist in Malaysia, highlighting the importance of understanding the role of ethnicity in Malaysian politics.

E. Urbanization

According to a 2022 report from the Department of Statistics Malaysia (DOSM), as of 2020, the urbanization rate in Malaysia was 75.1 percent, representing a 4.2 percent increase over the previous decade. The composition of this urban population is as follows: Bumiputera (62.6 percent), Chinese (28.6 percent), and Indians (8.1 percent), with 6.6 percent of this subset being over the age of 65. On the other hand, the rural population is composed of Bumiputera (90.4 percent), Chinese (6.3 percent) and Indians (2.4 percent) with 7.3 percent of this population subset being over the age of 65 [2].

The impact of urbanization on electoral outcomes in Malaysia has been widely studied and this section of the paper will examine existing findings on the issue, especially on the skewed composition of the rural and urban areas, attempts to drive urbanization in Malaysia, and the impact of urbanization on voters and politicians alike.

To understand urbanization in Malaysia, it is essential to examine the role of the New Economic Policy (NEP). Following the 1969 ethnic riots in Malaysia, NEP was implemented with the aims of eradicating poverty across all races and societal restructure to remove race-based identification with economic functions and geographic locations, with one underlying goal being to increase Bumiputera's participation in the urban sector [33].

Several issues have been identified with the NEP such as slowed economic growth and brain drain (as the non-bumiputeras seek countries with less discriminatory policies), however the policy achieved some success as Malaysia has reduced absolute income poverty, recorded increased levels of universal primary education [34], increased rates of urbanization, especially among the bumiputeras.

The NEP also led to the creation of a new Malay urban middle class who are more politically informed and more ideologically independent than the Malaysian rural inhabitants, focusing on political issues such as human rights and the rule of law. This situation may have altered voting patterns in Malaysia, leading to increased support for opposition parties and creating more space for political dissent in the country [35].

There is some dissent among researchers as to which of urbanization and ethnicity takes precedence in determining electoral outcomes in Malaysia. In their study on the most salient factors affecting the 2018 Malaysian 14th General Election, Ng et al. (2021) found that urbanization, followed by ethnicity, had the most significant influence on predicting the vote share of Barisan Nasional, the ruling party at the time of the election. They attributed the difference in political attitudes between urban and rural inhabitants in Malaysia to several reasons, including the difference in interests, whereby rural voters seek the fulfillment of immediate needs, such as food and state assistance, while urban voters seek higher-level needs, including quality of life and good governance. The metropolitan population also has better access to the internet, which has been utilized as a tool to foster political discourse and information sharing, enabling opposition parties to access it and garner support [36].

On the other hand, some authors caution against a binary analysis of both factors or not considering other factors when studying electoral outcomes in Malaysia. For example, Ong (2020) argues that urban or rural life highly correlates with different variables, including income and education, which makes it hard to pinpoint the specific variable that shapes political preferences. According to him, voters in rural electoral districts may not be living rural lives, and vice versa. Using OLS regression, Ong found that urban respondents to a survey about political attitudes consistently self-reported lower interests in politics compared to rural respondents and found no difference in the propensity of urban and rural voters to persuade others to participate in specific political activities [35].

Despite this dissent regarding the effects of urbanization on political attitudes, researchers have shared findings on how urbanization influences general attitudes, particularly among young people. Salleh (2013), in his study on the 1Malaysia concept, found that adolescents in urban areas were more likely to identify as Malaysians rather than as Malays, which was more common among rural groups. He suggests that this may be the result of what the urban youth see on television, read in books, and more importantly, from their interactions with friends. The rural youth, on the other hand, are not as exposed to media and rely on their family for information and their sense of identity; therefore, their likelihood of identifying more as Malays than Malaysians [37].

There are also differences in the governance issues that rural voters and urban voters are concerned about. According to Elangovan et al. (2022), rural voters were more concerned

about municipal issues such as infrastructural improvements and state assistance, while urban voters were more concerned about a cleaner (i.e., less corrupt) and stable government. Rural voters were aware of the bigger picture such as the economy and the future of democracy, but their concerns about the basic infrastructure they lacked took precedence over these big picture issues [38].

Politicians capitalize on the differences in rural and urban political attitudes in their messaging and campaigns. In their study on political advertising in Malaysia, Rahim et al. (2017) found that Barisan Nasional used messaging focused on everyday living issues, such as education, improved police force, and better services, which directly interest rural inhabitants, low-income earners, and the middle class. The opposition, on the other hand, focused on issues such as cronyism, corruption, and free education [39].

As Malaysia becomes increasingly urbanized and the younger generation replaces the older, it will be interesting to see the resultant change in political attitudes in rural and urban areas, and how that change reflects in electoral outcomes.

F. Age

In 2019, the Malaysian parliament lowered the voting age from 21 to 18, potentially increasing the population of new voters by about 16 percent and raising questions about who these new voters will benefit and the impact on future elections [40]. To swing these new voters to their side, Malaysian political parties have invested heavily in social media like TikTok and Twitter, which highlights the influence of this group of voters [41] but has led to the youth being concerned about the aggressiveness of the advertisements and campaigns [42].

The youth use tools such as the internet to engage in political discourse and activism but some researchers argue that the youth may have developed political apathy after their activism was undermined by lack of political change and the government through manipulation, spying, threatening etc [16], [39], [43]. Other researchers have also attributed the political apathy of the youth to a lack of opportunities to showcase political skills, their lack of knowledge, and excessive focus on entertainment and pleasure-seeking activities [44].

The youth rely heavily on social media for political information [42], are concerned about the state of the economy, and tend to vote outside of religious and ethnic sentiments [45]. A study by Abd Rahim et al. (2017) showed that the young voters were more attracted to political advertisements "that promote peace, stability, progress, pride and developments" [39]. This group tends to consider the party, its leaders' policies and attitudes, and their previous performance. It promises that, when making voting decisions, they believe that if they continue to vote, there might be some change in the political system, suggesting continued hope in democracy.

Following the recently concluded GE15 in 2022, there are opportunities for researchers to examine how the youth used social media during the lead-up to the elections and how their participation influenced the electoral process and outcome. It may also be useful for researchers to study how the use of TikTok and other platforms reflects the political attitudes of young people, especially in light of the recently lowered voting age. There also seems to be a disparity between the identified political attitude of the Malaysian youth, which is that of issue voters, versus the behavior of Malaysian TikTok users in the lead-up to GE15 where there were several incidents of ethnic and religious hate speech [46]. This raises questions about who these TikTok users are and what their agenda is.

G. Social Media in Malaysian Politics

The role of social media in Malaysian politics has been widely studied. The Internet Users Survey conducted by the Malaysian Communications and Multimedia Commission in 2020 reported an internet usage rate of 87.4 percent and smartphone usage of 98.7 percent. Political content sharing spiked in 2018 to 32.1 percent, an election year, but reduced to 17.2 percent in 2020, with the most popular sites being Facebook (91.7 percent), YouTube (80.6 percent) and Instagram (63.1 percent) (IUS 2020). Recently, TikTok has been gaining popularity in Malaysia. As of January 2023, Statista reported a penetration rate of 77.7 percent among adults over 18 years [47], with those between the ages of 19 and 25 years being the majority of TikTok users [48]. This usage rate is even more notable when the incidents prior to and during the recently concluded 15th General Elections, held in 2022, are considered. TikTok became a breeding ground for ethnic-based hate speech and violent extremism [49].

Scholars have employed the concept of affordances to examine how social media facilitates political discourse. Affordances such as sharing, likes, and timelines have increased information flow [50], reduced the time and effort required for participating in political discourse, improved access to political information, and also reduced the cost of collecting political information [51]. In authoritarian regimes like Malaysia, the availability of end-to-end encryption in social media applications such as WhatsApp and Telegram has also reduced the risks of government surveillance and retribution associated with participating in subversive political discourse [52].

In Malaysia, social media and its affordances have facilitated social mobilization movements such as the reformasi and the Bersih rallies. Social media has also played a significant role in electoral outcomes during elections. The reformasi movement, triggered by former deputy prime minister Anwar Ibrahim's detention, is acknowledged as the foundation for online political activism in Malaysia [28], [53], [54], even though it did not result in a regime change [55]–[57]. The Bersih rallies used social media and the internet to mobilize protesters, share information, and counter negative narratives, bypassing the restrictive traditional media [28], [58]. The rallies were able to induce some political changes in Malaysia, including the loss of the popular vote for the then-ruling coalition, Barisan Nasional, in the 13th general elections [59], [60].

The Malaysian government recognizes the influence of social media in driving political change. To maintain control

over the online space, the government has resorted to tactics including participating more actively on social media, using cybertroopers to spread government-endorsed narratives, creating bills aimed at punishing activists, and infiltrating activists' groups [61], [62].

Overall, social media has facilitated significant political change in Malaysia by allowing ease of coordination and connection [63]. However, some authors argue that social media does not necessarily change the underlying factors that affect politics in a society (Weiss 2013).

1) Government's Involvement: The Malaysian government has been accused of manipulating government agencies, controlling the news media, and using cybertroopers to sway public opinion. These measures have been aimed at protecting the government's interests and silencing critics, resulting in the introduction of laws that restrict free speech and freedom of expression, as well as the arrest of individuals who speak out against the government. The Sedition Act is one of the laws used to prosecute individuals who criticize the government. Overall, the government's engagement in Malaysia has been characterized by attempts to maintain its grip on power and silence those who oppose it.

2) Online Media Manipulation: After the racial riots of 13 May 1969, the media were placed under state control. From that time, the government and many in the media have portrayed the job of print and television media as primarily reporting "positive news" about government policies, racial harmony, and national identity [64]. The media's position as a "watch-dog" was viewed as a "Western" view of journalism and news reporting [65]. Malaysian authorities (and others in Asia) called for "developmental journalism", which is defined as journalism that aids in the process of nation-building and development and in which the press is not a natural foe of the government. Malaysian media practitioners have long advocated for more press freedom and a more independent, balanced media, but the BN has substantially undermined their efforts. Furthermore, the BN has frequently invoked 'developmental journalism' to deny the opposition access to national radio and television stations. While the Internet has increased online openness, voters in rural places face fewer different opinions due to a lack of Internet access. In Malaysia, media-freedom advocates and journalists have focused on the concept of 'maintaining a political balance' as a vital professional practice in journalism. That is, the ability to report on both the administration and the opposition, or to 'balance' favorable pro-government voices with critical ones. Following the GE13, some privately owned major media companies began to provide more balanced coverage [66].

3) Offline Media: To enhance sales and earnings, certain mainstream media businesses in Malaysia are promoting a business model of "balanced" coverage. This concept is intended to differentiate them from other companies in the media business, which are primarily pro-government in their coverage. For example, chief editor Abdul Jalil Ali from The daily Malay-language newspaper Sinar Harian said, "We believe if the newspaper is government-friendly, it won't be reader-friendly. Our readers determine our survival." Another example of this is the Oriental Daily. The Oriental Daily was another newspaper that promoted balance and truth during GE13. This Chinese-language newspaper was founded in 2003 by the KTS Group, which also owns a Sarawak forestry industry and has close ties to Abdul Taib Mahmud, who has been in power in Sarawak, Malaysian Borneo, for over 30 years. As a result, the Oriental Daily is not without political connections or ownership persuasions [67]. From 2006 to 2012, daily sales of most English and Bahasa-language newspapers fell, with only the tabloids The Sun and Harian Metro seeing an increase in circulation. The circulation of the New Straits Times in 2012 was 100,382, which was lower than that of any other major English, Malay, or Chinese-language daily newspaper in Malaysia (Malaysia Media Planning Guide 2013). While the drop in share price and newspaper readership is not unique to Malaysia, it is crucial to note that in more privately owned media, circulation revenue outnumbers advertising revenue. In several cases, it has been observed that Political leaders find a way to manipulate the content of the news. In some cases, they own the media house or they befriend the media-house owner to make the news content more government-friendly [67].

4) Electronic Media: The main television stations in Malaysia are all owned by the government-controlled Media Prima Berhad. However, Ananda Krishnan, Malaysia's second-richest person, controls Astro, a satellite-television operator with three million subscribers. Astro Awani, its news channel, was founded in late 2007, primarily to cover worldwide news. However, in 2008, it also covered the general election and revised its business strategy to hire more journalists to cover Malaysian news. Astro attempted to be more neutral in its coverage of the GE13 to increase its subscriber base and compete with the government-owned news channel TV3 [68].

5) Laws and Acts to suppress free speech: Since Malaysia gained independence in 1957, the Internal Security Act (ISA), the Official Secrets Act, the Sedition Act, and the Printing Presses and Publications Act have been the primary pieces of legislation that restrict press freedom and freedom of expression (PPPA). The ISA received the most criticism of all the government's strategies used to muzzle dissenting voices. The ISA was applied to opposition party and civil society leaders, most notably Anwar Ibrahim, the former Deputy Prime Minister, in 1998. It was also used to journalists and bloggers, such as Hishamuddin Rais, a columnist for Malaysiakini in 2001, and Raja Petra Kamaruddin, the most well-known blogger in the nation, in 2008. The Malaysian government amended the Printing Presses and Publications Act (PPPA) in April 2012 by removing the annual renewal requirement for licenses. The restriction on judicial review of any refusal, revocation, or suspension of a permit by the Home Minister was also lifted, and the publication (or other relevant person) was given the chance to be heard before a decision to revoke or suspend was made. Although that was a step forward, the reasons for suspension or revocation remain substantially the same. However, the law's provision allowing the public prosecutor to sentence journalists to up to three years in prison for disseminating "fake news" remained in

place [69].

6) Economy: The Mahathir administration's support for high-profile mega-projects as part of a larger objective to transform Malaysia into a developed economy by 2020 was one of its distinguishing features. The Multimedia Super Corridor's 1996 launch was the most notable of these initiatives. The initiative, now known as MSC Malaysia, sought to make Malaysia a regional IT hub by providing extensive tax advantages to foreign companies and significant government investment in high-speed broadband infrastructure in Cyberjaya, a newly constructed town and research park [70].

7) Influencing Government Agencies and Media: The government used several of their independent agencies to push their agendas. The Election Commission (EC), which was responsible for designating constituencies, maintaining an electoral role, and holding elections, was one of the bureaucratic institutions with the most direct influence on elections. It was ostensibly an independent organization, but it was part of the Prime Minister's Department, and all seven members were nominated by the prime minister, thus it actually advocated UMNO's interests (Ostwald 2017; see also Lim Teck Gee 2018). Tan Sri Abdul Rashid Abdul Rahman, the outgoing EC chairman, said in 2013 that he had ensured Malays remained in power through multiple elections.

Numerous legislative changes effectively eliminated the ability to appeal EC rulings, including a 2002 legislation that barred any judicial challenge to the electoral roll once it was gazetted [71]. The Registrar of Societies (ROS) was also crucial, as its clearance was required for political parties to compete in elections. ROS has a history of making the lives of opposing parties challenging. In the build-up to the 2013 elections, it kept everyone wondering until the very last minute before allowing the DAP to run under its flag, and it refused to recognize the opposition Pakatan Rakyat alliance as a party. Following Mahathir's dismissal of Malaysia's top judge in 1988, the judiciary likewise deferred to government wishes. Despite widespread skepticism, judges in several high-profile cases unanimously agreed with the government.

The two trials against Anwar Ibrahim are among the most well-known examples of judges making decisions that have been widely denounced by critics, including Malaysia's Bar Council and foreign legal experts [72], [73]. The BN kept and expanded its dominance over the major media during its leadership. The government primarily owned radio and television, and businesses typically held little private ownership with ties to UMNO. Organizations with ties to the BN controlled all the major print media outlets. On both print and television media sources, there were direct controls. Prime Minister Najib Razak, the United Malays National Organisation (UMNO), partially owned Utusan Malaysia, the largest Malay-language newspaper, which it exploited as a political instrument. Journalists, bloggers, and even a cartoonist who satirized Najib were arrested. 2015 saw Malaysia restrict media freedom when the Wall Street Journal published an article alleging that 700 million dollars from a public fund had gone into Najib's personal bank account. Additionally, it targeted the few publications that had resisted the crackdown, like MalaysiaKini and the Malaysia Insider [64].

H. Breaking Down Malaysian Politics

Analyzing the Motivations and Actions of Key Actors The Barisan Nasional (National Front) and the Pakatan Harapan are the two main coalitions in Malaysia's multi-party political system. Malaysian politicians frequently emphasize topics like corruption, ethnic and religious diversity, economic development, and human rights. Following their independence in 1957, the nation has undergone several political changes, including a change in administration in 2018 and a return to the previous ruling coalition in 2020. There are several political actors present in the Malaysian Political landscape:

1) United Malays National Organization (UMNO): The Barisan Nasional (BN) coalition, which includes UMNO as a founding member and the main dominant force, has been the main ruling force in Malaysia from the year of Malaya's independence in 1957 until its defeat in the general election of 2018. With a focus on race, UMNO seeks to defend Malay nationalist ideals and the concept of Ketuanan Melayu (Malay Supremacy), as well as the honor of the Malay people, Islam, and the nation as a whole. The party also seeks to uphold, defend, and promote Islam throughout Malaysia, as well as preserve Malay culture as the nation's cultural heritage [34].

2) Malaysian Chinese Association (MCA): One of the three original central com- ponent parties of the coalition party in Malaysia known as the Alliance Party, which later evolved into a larger coalition known as Barisan Nasional in Malay, or National Front in English, is the Malayan Chinese Association, a uni-racial political party that seeks to represent the Malaysian Chinese ethnicity. Since Malaysia gained its independence, MCA has had a considerable impact on the political landscape. The party was formerly Malaysia's largest Chinese political party, and it held sway from the early 1960s to the late 1960s.

3) Malaysian Indian Congress (MIC): A political organization in Malaysia is called the Malaysian Indian Congress (formerly Malayan Indian Congress). It is a founding member of the Barisan Nasional coalition, formerly known as the Alliance, which ruled the nation from its independence in 1957 until the 2018 elections. The party is one of the oldest in Malaysia and was among the first to struggle for Malayan Independence.

4) Malaysian Islamic Party/Parti Islam SeMalaysia (PAS): One of the significant actors and major Islamist political parties in Malaysia is called the Malaysian Islamic Party (Parti Islam Se-Malaysia). PAS's voting base is primarily located in Peninsular Malaysia's rural, conservative northern and eastern coasts, particularly in the states of Kelantan, Kedah, and others, as well as in some rural areas of Perak, due to the party's focus on Islamic extremism. The party was a member of the coalition that was in power at the time, Perikatan Nasional (PN), which was formed in response to the political unrest in Malaysia in 2020–21. In the states of Kelantan, Terengganu, Kedah, Perlis, and Sabah, the party is the only or one of the coalition partners in the government.

Table I: Coalitions, Member Parties, Ideologies, and Political Positions in Malaysia

Coalition	Member Parties	Ideology	Position
Barisan Nasional (BN)	United Malays National Organisation (UMNO)	• Ketuanan Melayu • National conservatism	Right Wing
	Malaysian Chinese Association (MCA)	• Malaysian Chinese interest • Social conservatism	Right Wing
	Malaysian Indian Congress (MIC)	• Malaysian Indian interest • Social conservatism	Right Wing
	Parti Bersatu Rakyat Sabah (PBRS)	• Sabah nationalism	Right Wing
Pakatan Harapan (PH)	People's Justice Party (PKR)	• Social liberalism • Malaysian reformism	Centre - Left
	Democratic Action Party (DAP)	• Social democracy • National secularism	Centre - Left
	National Trust Party (Parti Amanah Negara)	• Islamic modernism • National progressivism	Centre - Left
	United Progressive Kinabalu Organisation (UPKO)	• Sabah regionalism • Malaysian nationalism	Centre - Left
Perikatan Nasional (National Alliance)	Malaysian Islamic Party	• Islamism	Right-wing to Far-right
	Malaysian People's Movement Party	• Liberalism	Centre-left
	Malaysian United Indigenous Party	• Malay nationalism	Centre-right

5) Democratic Action Party (DAP): The Democratic Action Party (DAP) is a center-left social democratic political party in Malaysia. After defeating Barisan Nasional in the 2018 Malaysian general election and bringing an end to the party's 53-year stint in opposition, it joined the Pakatan Harapan (PH) coalition as one of the four constituent parties. Unfortunately, the coalition lost power after 22 months due to defections by its partner party, before it could complete its first term, which led to the political crisis in Malaysia in 2020. The PH coalition, of which the DAP was a part, was restored to power at the 2022 general election in Malaysia, albeit with a weaker majority, forcing it to form a unity government with opposition parties [34].

6) People's Justice Party (PKR): The National Justice Party and the socialist Malaysian People's Party were combined to form the reformist People's Justice Party, which was established on August 3, 2003. After former Deputy Prime Minister Anwar Ibrahim was arrested on April 4, 1999, during the height of the Reformasi movement, Wan Azizah Wan (wife of Anwar) Ismail founded the party's forerunner. The coalition of Pakatan Harapan (PH) includes the party as one of its principal members.

7) Parti Pribumi Bersatu Malaysia/Parti Pribumi Bersatu Malaysia (PPBM): A nationalist political party in Malaysia is called the Malaysian United Indigenous Party. The United Indigenous Association of Malaysia arrived before the party. Within the Perikatan Nasional alliance, it is a significant component party. From May 2020 through August 2021, the party held both the office of Prime Minister and the majority of the cabinet positions. In 2016, the United Malays National Organisation (UMNO) and the Barisan Nasional dissident organization Gabungan Ketua Cawangan Malaysia provided the party's founding members.

I. Chinese Influence in Malaysia

China's influence in Malaysia has been growing in recent years, and allegations have emerged that China has attempted to manipulate Malaysian domestic politics to serve its interests. An example of this is the alleged involvement of Chinese companies in corrupt practices to secure major infrastructure projects in Malaysia. Additionally, Chinese investments in Malaysia have been seen as a means for China to expand its influence in Southeast Asia. China has also been accused of using its economic power to influence Malaysian politics,

including by funding political parties and politicians who are seen as friendly and perceived to be pro-China. These actions have led to raised concerns in Malaysia about Malaysia's sovereignty and the possibility that China could use its influence to undermine Malaysian democracy.

1) Upheaval in Malaysian Politics: The Chinese government has attempted to influence the elections through both soft power, exploiting the apparent warmth of the Malaysian public, and more covert means of corruption, also known as sharp power. Beijing mainly fostered Chinese-Malaysians. The Malaysian Overseas Chinese Association (MCA), the overseas Chinese political party of the Najib government, stated that it had helped promote Chinese investment in Malaysia before the election. Party propaganda claims that "voting for the Najib coalition is synonymous with supporting China," and Chinese language media emphasize pro-China positions while refraining from criticizing Beijing for silencing opposition. These attempts were, at the very least, transparent.

However, in addition to more visible soft-power tools, China may have utilized covert and coercive means to bolster Najib. According to minutes from previously undisclosed meetings obtained by the Wall Street Journal, as Malaysia's election approached, Chinese officials told Najib that Beijing would pressure foreign countries to drop investigations into the graft-ridden, tanking 1Malaysia Development Berhad (1MDB) state fund, which was hemorrhaging money, and would even bail it out if the administration gave China stakes in Malaysian pipeline and rail projects. But unfortunately, China failed. On election day, Mahathir's coalition stunned Najib and his coalition to sweep into power [74]. The substantial Chinese FDI invested in Malaysia provides a significant chunk of the economic context for the political dilemma confronting Prime Minister Najib's administration. According to an announcement made by the Malaysian government in October 2016, the 620 km East Coast Rail Link (ECRL) would be built by China Communications Construction Co. and financed by China Export-Import Bank.

The ECRL, which will connect Kuantan Port on Peninsular Malaysia's east coast with Port Kuala Lumpurang on its west coast, will enable goods to be transshipped between the Malacca Strait and the South China Sea without requiring travel via Singapore [75]. As Mahathir has joined a political party that opposes the rule of current Prime Minister Najib Razak, many believe the dispute over Forest City is a sign that the general election that is anticipated for this year will revolve on Chinese Investment in Malaysia. Mustafa Izzuddin, for instance, forecasts that "Mahathir and other Malay politicians from the anti-Najib camp will use the sheer Chinese investments into Malaysia to criticize Najib as selling Malaysia's internal sovereignty to China to the extent of drifting into the China orbit and becoming its satellite state." China Railway Group, a Chinese state-owned enterprise, has formed a consortium with a local developer to purchase 60 percent of the land parcel for the 160 billion Bandar Malaysia real estate project, located at the Kuala Lumpur terminus of the Singapore-Kuala Lumpur high speed rail line, from 1Malaysia Development Berhad - the Malaysian state-owned investment fund whose financial troubles, as we shall shortly see, have triggered the current political crisis.

2) The Debate Over Chinese Investment and National Sovereignty in Malaysia: A dangerously racist political issue in Malaysia has attracted China because of a public disagreement between the Sultan of Johor and former Malaysian Prime Minister Mahathir Mohamad. The dispute relates to Forest City, a huge 100 billion US dollars real estate undertaking being built in Johor Bahru, the state capital of southern Malaysia's Johor. This property is being developed by Country Garden Pacific View, a joint venture between the Chinese developer Country Garden and a local developer wholly owned by the Sultan of Johor.

The Sultan of Johor argued that Mahathir "twisted" the facts to incite "fear, using race, to fulfill his political motives" in response to his accusations that the 700,000 future residents of Forest City will be mainland Chinese nationals, "that citizenships will be given away, and that vast tracts of land have been sold to the Chinese." The Sultan stated that the project, which is being built on reclaimed land, will instead "increase Johor land size and sovereignty" and that the residential units are "not just for Chinese investors, but for anyone around the world, including Johoreans," in response to Mahathir's accusations that "Johor is surrendering land to the Chinese and that we are giving up our sovereignty." For- est City, which is being constructed on four artificial islands, will eventually "home 700,000 people on an area four times the size of New York's Central Park," as well as "office towers, parks, hotels, shopping malls, and an international school." It is one of nearly 60 real estate developments in Johor's Iskandar Malaysia special economic zone — an area three times the size of Singapore — that has attracted considerable FDI from Chinese developers such as Country Garden (Lim, 2016; Mahrotri and Choong, 2016). The Sultan also noted that while Mahathir was Prime Minister, he encouraged Malaysians to "look East," but now he criticizes Chinese capitalists who come to invest in Malaysia. The Chinese embassy in Malaysia stated emphatically that "someone welcomed Sino-Malaysian cooperation while in power but stoked the flames of anti-Chinese sentiment after... Claiming that Chinese investment is snatching Malaysian jobs is a total lie with a hidden objective" [76], [77].

3) Significant Chinese Investment: The massive Chinese FDI invested in Malaysia constitutes a significant element of the economic backdrop to Prime Minister Najib's leadership challenge. The Malaysian government announced in October 2016 that China Communications Construction Co. will build, and China Export-Import Bank will fund the 620 km East Coast Rail Link (ECRL), which is expected to cost 55 billion. The ECRL will connect Kuantan Port on Peninsular Malaysia's east coast with Port Kuala Lumpurang on the west coast, providing a land bridge that will allow products to be transshipped between the Malacca Straits and the South China Sea without passing through Singapore.

The Malaysia-China Kuantan Industrial Park (MCKIP), which has drawn investment in "high-tech businesses including stainless steel goods, electrical and electronics, information

communication technology, and renewable energy," would profit from this increase in transportation alternatives. The MCKIP has been paired with the China-Malaysia Qinzhou Industrial Park in the Chinese province of Guangxi, and the ongoing development and modernization of Kuantan Port, in which China's Guangxi Beibu Gulf International Port Group has a 40 percent stake, will help to increase trade between Malaysia and China [75].

A 30 billion joint venture has been established to build the Melaka Gateway in the state of Melaka on the west coast of Peninsular Malaysia by local developer KAJ Development and the Power China International Group of China. A natural island will be "marked as a container and bulk terminal, shipbuilding and ship repair services, and a maritime industrial park" in addition to three artificial islands that will be used for "different tourism, commercial, property, and marine activities" [78]. The project would also involve building the Melaka Gateway Port, which is expected to be finished in 2019. It will cost 8 billion.

Chinese rail companies are vying for the Singapore-Kuala Lumpur high-speed rail project. Although the contract to build the high-speed rail link between Singapore and Kuala Lumpur has not yet been released, "local media reports imply that Singapore prefers a Japanese or European bidder while Malaysia favors a Chinese one" (Martin, 2016). The Malaysian state- owned investment fund 1Malaysia Development Berhad (1MDB), whose financial issues, as we shall see in a moment, have led to the current political crisis, has in fact formed a consortium with a local developer to purchase 60 percent of the land parcel for the 160 billion Bandar Malaysia real estate project, located at the Kuala Lumpur terminus of the Singapore-Kuala Lumpur high speed rail line.

4) South-China Sea Dispute: The recent increase in Chinese FDI appears to have softened the Malaysian government's former firm attitude against China in the case of the South China Sea dispute. According to Urchick (2017), Prime Minister Najib traveled to Beijing after newly-elected Philippine President Rodrigo Duterte took steps to normalize relations with Beijing "to meet with China's leadership and agree to negotiate their dispute bilaterally, which resulted in the signing of 34 billion dollars in trade deals and a naval vessel arms sale to Malaysia," Urchick (2017) recollects. Manila and Kuala Lumpur undoubtedly won't want to kill the goose that lays the golden eggs given the scope of Chinese generosity. Hence, "the Philippines and Malaysia will not retract or drop their claims in the South China Sea but will instead work to keep their dissatisfaction quiet for the sake of bilateral trade deals and will continue with their projected military purchases." Chinese Investment does not, however, provide Beijing free reign to act without anticipating a reaction from its allies. For instance, the satellite photographs showing the most recent Chinese installation of military hardware in the Spratly Islands prompted the governments of Malaysia and the Philippines to ask Beijing for clarifications. Even though they are inadequate, their actions demonstrate to their citizens that their governments have not allowed China to militarize the disputed islands [79].

5) Chinese support for 1MDB Project: The financial assistance for 1MDB came from some Chinese FDI in Malaysia. Chinese state-owned companies have achieved this through the purchase of 1MDB assets, such as China General Nuclear Power Co.'s 17 billion acquisition of Edra Global Energy. Such Chinese state-owned companies' acquisitions of Malaysian infrastructure assets have upset opposition lawmakers, including Democratic Action Party member Ong Kian Ming, who asserted that: "The Malaysian government needs to be fully transparent on the specifics of this deal. The Malaysian taxpayer deserves to be informed of the full cost of the rescue because nothing is ever free. Most recently, China generously gave 1MDB finance in exchange for state assets to assist 1MDB in paying off its debt to Abu Dhabi's state-owned International Petroleum Investment Co., which totaled about USD 6.5 billion. Although Chinese Investment has assisted 1MDB in surviving its financial difficulties, doing so has put China at political danger because 1MDB's issues are related to a still-developing international corruption scandal in which Prime Minister Najib has a significant stake [80]. After the US Justice Department launched civil proceedings in July 2016, implicating Najib in the scam, the Malaysian government shifted its stance from the US to China. As a result, Najib made the aforementioned crucial trip to Beijing. Hence, China has become a crucial component of the opposition coalition's criticism of Najib's leadership [76].

6) Potential Interference and Radicalization in Malaysian Domestic Politics: Given that the political conflict is radicalized, the involvement of China is particularly risky. The radicalization first became apparent in 2015, when it was discovered that Najib's personal bank accounts contained suspicious deposits totaling around USD 700 million. Following large-scale opposition demonstrations demanding the removal of Najib, pro-government demonstrations were held, where the conflict was reframed as an effort by Malaysia's Chinese minority to restrict the privileges enjoyed by the Malay majority. This radicalization has dangerously increased the ethnic tension in the political conflict because Malaysia has a history of ethnic rioting [81]. The Chinese administration has had to be cautious in its interactions with the Malaysian government cause of the racist aspect of the political disagreement. After publicly stating that "China, Malaysia's top trading partner, would not hesitate to speak out against any threat that may affect the country's ties with Malaysia and that Beijing is opposed to discrimination against races and any form of extremism, the Chinese ambassador to Malaysia was summoned to Malaysia's Foreign Ministry in September 2015. The ambassador's remarks came after pro-government Malay groups threatened racial violence following an anti-government event that Malaysians of Chinese ancestry predominantly attended. The Chinese embassy clarified that the ambassador's message was only an "act of goodwill" and an expression of "the desire that Malaysia stays united, successful, and harmonious," whereas the Malaysian government viewed it as meddling in Malaysian domestic matters [82]. The ambassador expressed hope that Malaysia will be able to "keep national unity and stability and

ethnic harmony," while the Chinese Foreign Ministry clarified that China does not "interfere in other countries' domestic politics or intervene in other countries' internal affairs".

III. Bibliographic Analysis

This section provides an analytical overview of the literature referenced in this study, presented through two key visual methods: a citation network diagram and a temporal timeline graph. These tools help to map the intellectual landscape and historical progression of scholarship related to Malaysian politics and digital influence.

1) Literature Network Diagram: The citation network diagram (Figure 5) reveals the relationships between selected literature, represented by green nodes, and the broader body of related literature, represented by blue nodes. The arrows indicate the direction of citation, while their thickness reflects the level of scholarly influence or frequency of reference. The structure of this network reveals that certain scholars, notably Abbott (2001, 2013, 2015) and George (2005), occupy a central and influential role in shaping the discourse surrounding media control, government narratives, and democratic transitions in Malaysia. Their work forms a foundation upon which more recent studies have built. The diagram also reveals distinct clusters that align with the thematic structure of this paper, such as media censorship, ethnic voting patterns, online mobilization, and foreign interference, notably from China.

Additionally, the presence of bridging authors such as Postill (2014) and Khoo (2016) connects earlier foundational research with emerging perspectives on digital activism, indicating a continuous evolution of the scholarly conversation. These authors serve as intellectual conduits, drawing together otherwise disconnected bodies of work, particularly linking traditional political theory with recent studies on internet-enabled political engagement and reform movements.

2) Timeline Diagram: Complementing the network diagram, the timeline graph (Figure 4) visually depicts the chronological distribution of the cited literature. This visualization demonstrates a clear shift in scholarly focus over time. Before 2010, research primarily emphasized authoritarian state control, traditional ethnic politics, and media regulation. Moving into the 2010s, there is a noticeable expansion into digital domains, reflecting a growing awareness of the transformative power of social media in Malaysian civic life. This trend intensifies after 2016, coinciding with political shifts during and after the 14th General Election (GE14), which brought regime change and intensified public discourse online. Recent years have seen an increase in literature addressing platform-specific behaviors, especially the use of TikTok and WhatsApp for political messaging, as well as concerns over hate speech and misinformation.

Together, these visual analyses support the paper's broader argument that Malaysian political discourse has undergone a significant transformation, heavily influenced by the interplay of technology, identity, and governance. The scholarly record reveals a dynamic and expanding field of inquiry, one that continues to evolve in tandem with Malaysia's shifting sociopolitical landscape.

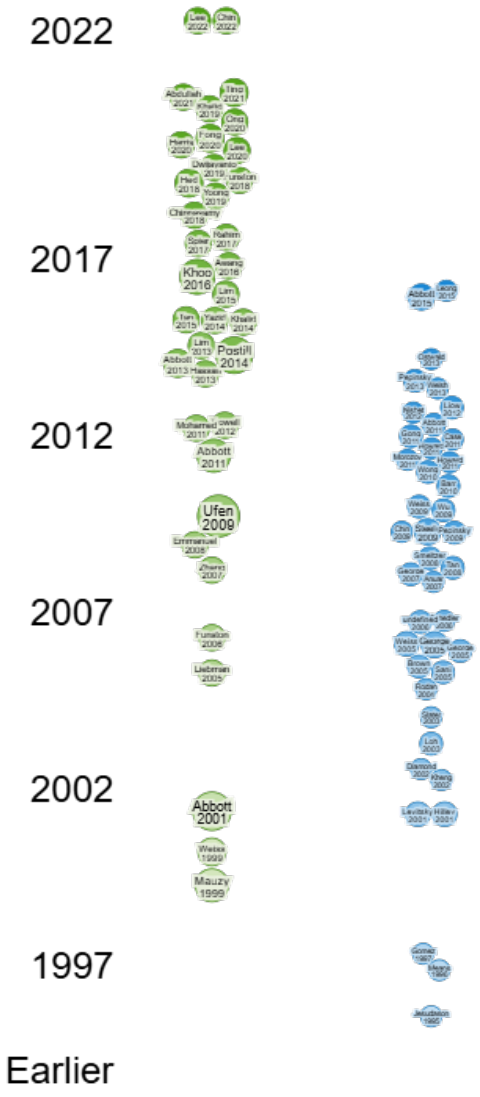


Figure 4: Timeline Graph of Selected (Green) and Related (Blue) Literature published

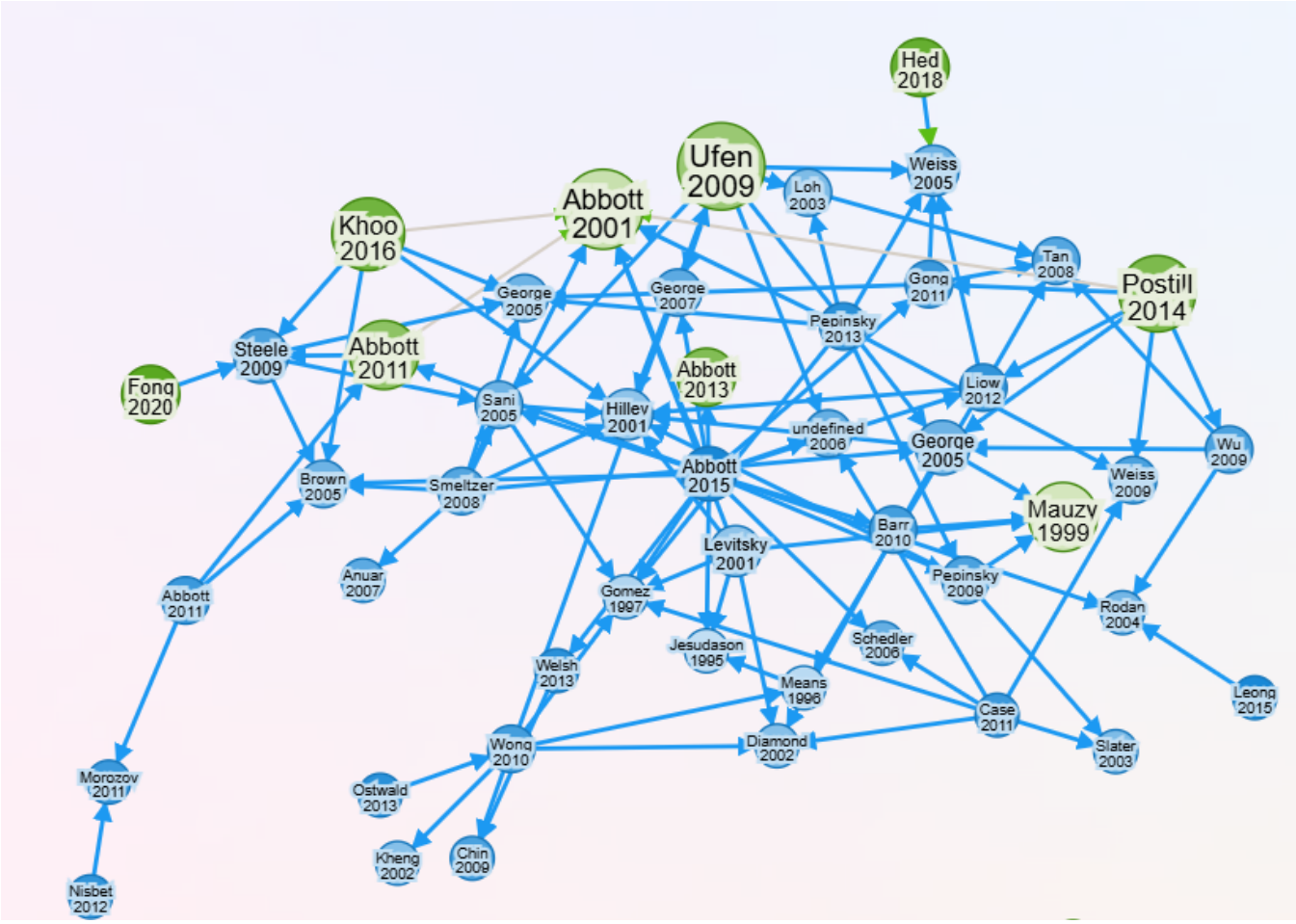


Figure 5: Citation network diagram of selected (green) and related (blue) literature.

IV. Network Analysis

Building upon the bibliographic analysis of Malaysian political communication studies, this section examines a specific methodological advancement in understanding digital political discourse through social network analysis. To further investigate the bibliographic and thematic insights identified in our survey of Malaysian political discourse, we conducted a computational analysis of the 2022 Malaysian General Election (GE15), with a specific focus on Instagram as a digital space for political engagement. We analyzed 53,116 Instagram posts collected using election-related hashtags—including #UndiHarapan, #KitaBoleh, #PH, #Election2022, #Malaysia, #AnwarIbrahim, #PakatanHarapan, #GE15, #KelasDemokrasi, #PRU15, and #MalaysiaMemilih—using the APIFY scraper tool. Leveraging this dataset, we employed a two-step methodological framework combining Contextual Focal Structure Analysis (CFSA) and topic modeling.

1) Contextual Focal Structure Analysis (CFSA): The CFSA methodology represents a significant advancement over traditional social network analysis by incorporating contextual information alongside user interactions, enabling a more nuanced understanding of network dynamics. Unlike simpler models that focus solely on user-user connections, CFSA integrates multiple layers of information, such as shared topics or hashtags, to provide a richer representation of social interactions.

The analytical process involved three key phases:

- **Network Construction:** The researchers generated a co-occurrence network of users based on mentions, where users who mentioned each other in their posts were considered linked, creating a web of interactions that reflected the discourse around the Malaysian general election.
- **CFSA Implementation:** The modified CFSA model accepted users and the links between them based on mentions, representing a coupling matrix. The outcome included the smallest possible contextual focal structure sets, comprising influential users within different communities who were frequently mentioned or who mentioned others in election-related posts.
- **Topic Analysis:** The researchers employed Latent Dirichlet Allocation (LDA) topic modeling to extract and analyze the most frequent issues from textual content associated with each focal structure. The model achieved a perplexity score of 8.07 and a coherence value of 0.59, demonstrating the quality and reliability of the topic

modeling results.

CFSA enabled us to map the underlying interaction network of Instagram users who were actively shaping the electoral conversation. This technique identified sets of influential actors not just by popularity but by their relational embeddedness within election-related discourse. Our network was built from user mentions, constructing an interaction matrix that captured how individuals and entities engaged with each other during the campaign period.

Through this, we revealed tightly connected clusters of users—primarily journalists, media houses, and political party affiliates—who collectively functioned as key drivers of online narrative framing.

In parallel (Figure 6), we employed Latent Dirichlet Allocation (LDA) for topic modeling, extracting coherent themes from the text content of the posts. The most prominent themes centered around voter empowerment, media framing, leadership identity (notably Anwar Ibrahim), and coalition branding (e.g., “KamiAWANI”, “Malaysia Memilih”). The network visualization underscored the convergence of media and political communication, indicating a blurring of boundaries between news outlets and partisan actors. Interestingly, while media figures appeared to span multiple narrative clusters, certain political actors and party-linked profiles dominated isolated sub-networks, suggesting varied levels of online influence and cross-sector collaboration.

A. Impact & Implication

This analysis makes several significant contributions to understanding Malaysian political scenarios, beginning with its methodological innovation in political discourse analysis. The application of CFSA to political discourse analysis provides a more sophisticated approach than traditional network analysis methods, demonstrating how the combination of network analysis with content analysis can effectively decode the multifaceted nature of online political communication. This methodological advancement makes a substantial contribution to the broader field of digital democracy studies by providing researchers with a robust framework for analyzing complex political networks that extends beyond simple connectivity patterns to incorporate contextual relationships and thematic coherence.

The findings reveal a concerning trend in media-politics convergence, where the close interconnection between media professionals and politicians underscores a blurring of lines between these traditionally distinct sectors. This convergence raises critical questions about media independence and the framing of political narratives, as the symbiotic relationship between media coverage and political messaging appears to influence how the public perceives and engages with election-related information. The implications extend beyond mere professional relationships to fundamental concerns about information flow and potential bias in political coverage, suggesting that the traditional role of media as an independent watchdog may be compromised in Malaysia’s digital political landscape.

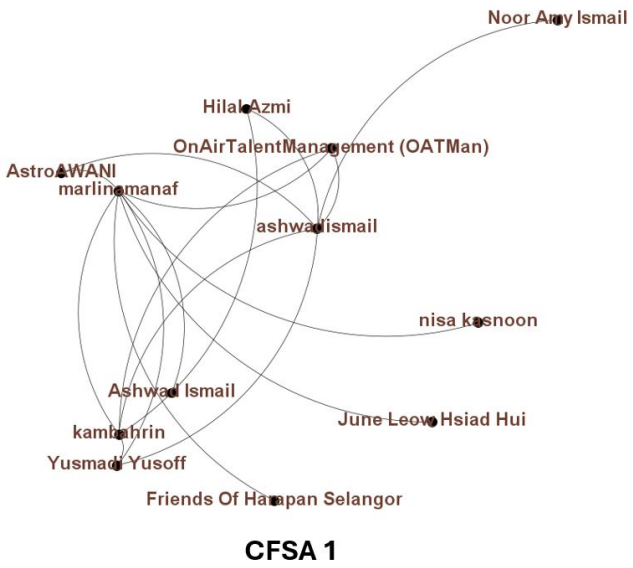
These methodological insights seamlessly connect to broader implications for digital democracy, as the study highlights the significant role of social media, particularly Instagram, in modern political communication. The identified focal structures reveal an intricate interplay between journalists, media houses, politicians, and political parties, demonstrating how digital platforms have become crucial for political engagement and information dissemination. This transformation represents a fundamental shift in how political discourse is constructed and consumed, moving from traditional gatekeeping models to more complex, interconnected networks where influence flows through multiple channels simultaneously.

However, the research also reveals troubling concerns about political representation that arise from this digital transformation. The explicit mention of “Friends Of Harapan Selangor" without strong representation from other political parties indicates a potential disparity in online presence or influence among different political factions during this crucial period, suggesting unbalanced digital political discourse. This imbalance extends beyond simple visibility to questions of democratic representation in digital spaces, where certain political voices may dominate while others remain marginalized or underrepresented in the online conversation.

The analysis of information flow patterns further illuminates these concerns, as the network structure not only highlights the central role of media in shaping public opinion during elections but also points to potential challenges in maintaining a balanced and diverse political discourse in the Malaysian online sphere. The concentration of influence within specific focal structures suggests that information dissemination may be controlled by relatively small networks of interconnected actors, potentially limiting the diversity of perspectives and narratives available to the public during critical democratic processes.

This research expands the bibliographic understanding of Malaysian political communication by providing empirical evidence on how digital platforms reshape the structures of political discourse, moving beyond theoretical frameworks to demonstrate concrete patterns of influence and interaction. The CFSA methodology offers future researchers a robust framework for analyzing complex political networks. At the same time, the findings shed light on the evolving dynamics of Malaysian democracy in the digital age. The study’s revelation of media-political convergence patterns provides crucial insights for understanding contemporary challenges to media independence and democratic discourse in Malaysia’s increasingly digital political landscape, suggesting that traditional concepts of media autonomy and political representation require reconceptualization in the context of social media-driven political communication.

Ultimately, this integrated methodological approach revealed how digital platforms, such as Instagram, have evolved into key arenas for political discourse in Malaysia, where influence is exerted not just through traditional metrics like post volume or likes, but also through strategic positioning within networked conversations. Our findings illustrate the dynamic



Frequent Keywords in Topics
marlinamanaf, luqmanhariz, Malaysia chooses, KamiAWANI, MalaysiaMemilih, Anwar Ibrahim

Breakdown of CFSA 1	
Journalist/Media House	Hilal Azmi, Kambahrin, Ashwad Ismail, Marlinamanaf, Nisa Kasnoon, Astro AWANI, OnAirTalentManagement (OATMan)
Politician/Political Party	June Leow Hsiad Hui, Friends Of Harapan Selangor, Yusmadi Yusoff

Figure 6: CFSA 1 set (among 11 focal structures) with breakdown and Frequent Keywords obtained by LDA

interplay between media narratives, public engagement, and political communication strategies in a hybrid media system increasingly shaped by social platforms.

V. Conclusion

Malaysia is a multicultural nation with a complex political system that continues to evolve in response to diverse social, economic, and technological forces. In this survey paper, we examined the various factors shaping the political landscape in Malaysia through a comprehensive approach that combines bibliographic analysis with advanced social network analysis. Our investigation encompassed the political attitudes of different voter subgroups, the role of social media, government actions, and the impact of Malaysia's relationship with China. It introduced empirical evidence through the Contextual Focal Structure Analysis of the 2022 Malaysian general election.

Through our bibliographic analysis, we identified and examined how political attitudes are formed across ethnicities, gender and age, shedding light on how voting choices are made across these demographic segments. We highlighted the evolution and growing importance of social media as a tool for political discourse and mobilization, demonstrating how these platforms have given citizens the tools to criticize government failings while simultaneously allowing the government to further push their propaganda. Our examination of government actions revealed several incidents where direct political interference has obstructed democratic processes in the country. Our analysis of the Chinese factor provided detailed insights into how Malaysia's relationship with China influences domestic political dynamics.

Building upon this foundational understanding, our application of Contextual Focal Structure Analysis to Instagram data from the 2022 Malaysian general election provided empirical validation of social media's transformative role in political communication. The CFSA methodology revealed complex networks of influence involving journalists, media houses, politicians, and political parties, demonstrating how digital platforms have fundamentally reshaped the structures of political discourse. Our findings uncovered concerning patterns of media-politics convergence, where traditional boundaries between journalism and political advocacy have become increasingly blurred, raising critical questions about media independence and democratic representation in Malaysia's digital age.

The network analysis revealed significant disparities in online political representation, with certain political factions dominating digital discourse while others remained marginalized. This empirical evidence corroborates our bibliographic findings regarding the uneven distribution of political power and influence in Malaysia's contemporary landscape. The identification of key focal structures and information flow patterns provides concrete evidence of how political narratives are constructed and disseminated through interconnected networks of media and political actors, offering a more nuanced understanding of influence dynamics than traditional analysis methods.

Overall, this survey paper highlights the complex and evolving nature of Malaysian politics, offering both theoretical background and empirical evidence to inform future research in the area. The combination of a comprehensive literature review with advanced social network analysis demonstrates the value of integrating multiple methodological approaches to understand contemporary political phenomena. The authors hope to expand this research framework by conducting longitudinal social media analytics across various platforms and elections, while developing more sophisticated methods for detecting and analyzing political influence networks. Future research will focus on examining the real-world impact of digital political discourse on voting behavior and democratic outcomes, contributing to a deeper understanding of how

digital transformation is reshaping Malaysian democracy and political participation.

Acknowledgments

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Army Research Office (W911NF-23-1-0011, W911NF-24-1-0078, W911NF-25-1-0147), U.S. Office of Naval Research (N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Defense Advanced Research Projects Agency, the Australian Department of Defense Strategic Policy Grants Program, Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment, and the Donaghey Foundation at the University of Arkansas at Little Rock. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

References

[1] N. Yousefi, N. Agarwal, K. Watts DiCicco, and M. S. Morshed, "Examining the impact of toxicity on community structure in social networks", in *SOTICS 2024: The Fourteenth International Conference on Social Media Technologies, Communication, and Informatics*, Venice, Italy: International Academy, Research, and Industry Association (IARIA), 2024, pp. 17–22, ISBN: 978-1-68558-198-5.

[2] Department of Statistics Malaysia, *Key findings of population and housing census of MALAYSIA 2020: Urban and rural [press release]*, December 23, 2022.

[3] Prime Minister's Department, Department of Statistics Malaysia, *Statistics on women empowerment in selected domains, Malaysia, 2021 [press release]*, November 17, 2021.

[4] World Bank, *Malaysia - World Bank Gender Data Portal*.

[5] K. A. T. Khalid, "Women and politics: Social construction and a policy of deconstruction", *Journal of Social Sciences*, vol. 10, no. 3, pp. 104–113, 2014. DOI: 10.3844/jssp.2014.104.113.

[6] M. Mohamad, "Getting more women into politics under one-party dominance: Collaboration, clientelism, and coalition building in the determination of women's representation in Malaysia", *Southeast Asian Studies*, vol. 7, no. 3, pp. 415–447, 2018. DOI: 10.20495/seas.7.3_415.

[7] M. Stivens, "Becoming modern in Malaysia: Women at the end of the twentieth century", in *Routledge eBooks*, 2020, pp. 16–38. DOI: 10.4324/9781003118411-2.

[8] Wan Azizah, "Women in politics: Reflections from Malaysia", in *International IDEA*, 2001, pp. 191–202.

[9] N. H. Selamat and N. Endut, "Bargaining with Patriarchy and entrepreneurship: Narratives of Malay Muslim women entrepreneurs in Malaysia", *Kajian Malaysia*, vol. 38, no. Supp.1, pp. 11–31, 2020. DOI: 10.21315/km2020.38.s1.2.

[10] U. Mellström, "The intersection of gender, race and cultural boundaries, or why is computer science in Malaysia dominated by women?", *Social Studies of Science*, vol. 39, no. 6, pp. 885–907, 2009. DOI: 10.1177/0306312709334636.

[11] N. M. Noor, "Roles and women's well-being: Some preliminary findings from Malaysia", *Sex Roles*, vol. 41, no. 3–4, pp. 123–143, 1999.

[12] S. M. K. Aljunied, "Against multiple hegemonies: Radical Malay women in colonial Malaya", *Journal of Social History*, 2013. DOI: 10.1093/jsh/sht056.

[13] M. Mohamad, "The politics of gender, ethnicity, and democratization in Malaysia: Shifting interests and identities", in *Oxford University PressOxford eBooks*, 2002, pp. 347–383. DOI: 10.1093/0199256454.003.0011.

[14] J. Herlijanto, "Between ethnicity and gender: Chinese women in contemporary Malaysian politics", in *Asian Scholarship Foundation. Sixth Annual Conference*, Bangkok, Thailand.

[15] U. A. A. Zakuan, M. A. M. Sani, N. Abdullah, and Z. Azmi, "How did we Choose?: Understanding the Northern female voting behaviour in Malaysia in the 14th general election", *Intellectual Discourse*, vol. 26, no. 2, pp. 859–882, 2018.

[16] S.-H. Ting and S. S. S. Ahmad, "Everyday interactions and political participation of Malaysian youth", *Journal of Youth Studies*, vol. 25, no. 5, pp. 616–635, 2021. DOI: 10.1080/13676261.2021.1923672.

[17] P. J. Yeong, "How women matter: Gender representation in Malaysia's 14th general election", *The Round Table: The Commonwealth Journal of International Affairs*, vol. 107, no. 6, pp. 771–786, 2018. DOI: 10.1080/00358533.2018.1545943.

[18] C. H. M. Ng, *The Hazy New Dawn: Democracy, Women and Politics in Malaysia*, https://doi.org/10.2139/ssrn.1716586, Social Science Research Network, 2010. DOI: 10.2139/ssrn.1716586.

[19] M. N. M. Yazid, "Colonial policy and the impact to the politico-economy stability after independence: The case of Indonesia under the Dutch and Malaysia under the British", *Review of History and Political Science*, vol. 2, no. 3 & 4, 2014. DOI: 10.15640/rhps.v2n3-4a4.

[20] K. Ostwald and S. E. Oliver, *Four arenas: Malaysia's 2018 election, reform, and democratization*, APSA Pre-Print, 2019. DOI: 10.33774/apsa-2019-6lprv.

[21] P. C. Phan, *Christianities in Asia*. John Wiley & Sons, 2010.

[22] R. S. Milne and D. K. Mauzy, *Malaysian Politics Under Mahathir*. Routledge, 1999.

[23] A. R. Moten, "Changing political culture and electoral behavior in Malaysia", *Asian Affairs: An American Review*, vol. 38, no. 1, pp. 39–56, 2011. DOI: 10.1080/00927678.2010.548201.

[24] L. W. Pye and M. W. Pye, *Asian Power and Politics: The Cultural Dimensions of Authority*. Harvard University Press, 2009.

[25] J. R. Kennedy, "Leadership in Malaysia: Traditional values, international outlook", *Academy of Management Perspectives*, vol. 16, no. 3, pp. 15–26, 2002. DOI: 10.5465/ame.2002.8540292.

[26] M. A. Khalid and L. Yang, "Income inequality and ethnic cleavages in Malaysia: Evidence from distributional national accounts (1984–2014)", *Journal of Asian Economics*, 2021. DOI: 10.1016/j.asieco.2020.101252.

[27] S. Leong, "The hindraf saga: Media and citizenship in Malaysia", in *Communication, Creativity and Global Citizenship: Refereed Proceedings of the Australian and New Zealand Communications Association Annual Conference*, 2009.

[28] L. Yangyue, "Controlling cyberspace in Malaysia", *Asian Survey*, vol. 54, no. 4, pp. 801–823, 2014. DOI: 10.1525/as.2014.54.4.801.

[29] A. A. Awang, S. S. S. Sheikh, A. M. Lokman, and A. B. Saifuddin, "A theoretical analysis of racial integration through new economic policy and 1Malaysia concept", *Advanced Science Letters*, 2016. DOI: 10.1166/asl.2016.6614.

[30] N. M. Noor and C. Leong, "Multiculturalism in Malaysia and Singapore: Contesting models", *International Journal of Intercultural Relations*, vol. 37, no. 6, pp. 714–726, 2013. DOI: 10.1016/j.ijintrel.2013.09.009.

[31] A. Harris and A. Han, "1Malaysia? young people and everyday multiculturalism in multiracialized Malaysia", *Ethnic*

and Racial Studies, vol. 43, no. 5, pp. 816–834, 2020. DOI: 10.1080/01419870.2019.1580379.

[32] PUSAT KOMAS Malaysia, *Malaysia racism report 2022*, Retrieved April 25, 2023, 2023.

[33] N. Hassan, N. I. Jaafar, R. N. R. Ariffin, A. A. Samah, and M. S. Jaafar, "Perceptions on quality of life in Malaysia: The urban-rural divide", *Planning Malaysia Journal*, vol. 11, no. 3, 2013. DOI: 10.21837/pm.v11i3.106.

[34] K. Rajandran and C. Lee, "Politics in Malaysia: A discourse perspective", in *Asia in Transition*, Springer Nature, 2023, pp. 1–15. DOI: 10.1007/978-981-19-5334-7_1.

[35] E. Ong, "Urban versus rural voters in Malaysia: More similarities than differences", *Contemporary Southeast Asia: A Journal of International and Strategic Affairs*, vol. 42, no. 1, pp. 28–57, 2020. DOI: 10.1355/cs42-1b.

[36] J. W. J. Ng, G. J. Rangel, and E. P. Y. Chin, "Did urbanization or ethnicity matter more in Malaysia's 14th general election?", *Contemporary Southeast Asia*, vol. 43, no. 3, pp. 461–495, 2021. DOI: 10.1355/cs43-3b.

[37] S. M. Salleh, "Unity in diversity", *Journal of Asian Pacific Communication*, vol. 23, no. 2, pp. 183–195, 2013. DOI: 10.1075/japc.23.2.01moh.

[38] N. Elangovan, J. Ong, and T. Zalizan, *Malaysia ge2022: The rural-urban divide — what the different groups of voters are looking for*, November 15, 2022.

[39] M. H. A. Rahim, N. Lyndon, and N. S. P. Mohamed, "Transforming political advertising in Malaysia: Strategizing political advertisements towards first-time and young voters in Malaysian ge 14", *Jurnal Komunikasi: Malaysian Journal of Communication*, 2017. DOI: 10.17576/jkmjc-2017-3301-23.

[40] C. Lee, *The lowered voting age in Malaysia: Who will benefit?*, ISEAS YUSOF ISHAK INSTITUTE, Issue 5, 2020.

[41] J. Chin, "Racism towards the Chinese minority in Malaysia: Political Islam and institutional barriers", *The Political Quarterly*, vol. 93, no. 3, pp. 451–459, 2022. DOI: 10.1111/1467-923x.13145.

[42] S. Chinnasamy and N. M. Azmi, "Malaysian 14th general election: Young voters & rising political participation", *The Journal of Social Sciences Research*, vol. SPI4, pp. 125–138, 2018. DOI: 10.32861/jssr.spi4.125.138.

[43] A. Johns and N. Cheong, "Feeling the chill: Bersih 2.0, state censorship, and "networked affect" on Malaysian social media 2012–2018", *Social Media and Society*, vol. 5, no. 2, 2019. DOI: 10.1177/2056305118821801.

[44] N. H. Abdullah, I. Hassan, M. F. Bin Ahmad, N. A. Hassan, and M. M. Ismail, "Social media, youths and political participation in Malaysia: A review of literature", *International Journal of Academic Research in Business and Social Sciences*, vol. 11, no. 4, pp. 845–857, 2021, n.d.

[45] A. Dwijayanto, Y. U. Afif, and K. Fathoni, "Managing democracy in Malaysia (identity, minorities, and representation)", *Jurnal Aristo (Social, Politic, Humaniora)*, vol. 08, no. 1, pp. 173–191, 2020.

[46] N. Jalli, *How tiktok became a breeding ground for hate speech in the latest Malaysia general election*, March 23, 2023.

[47] Statista, *Tiktok penetration in selected countries and territories 2023*, March 13, 2023.

[48] Statista, *Share of tiktok users in Malaysia 2022, by age group*, April 24, 2023.

[49] R. Latiff, *Tiktok on "High Alert" in Malaysia as tensions rise over election wrangle*, November 23, 2022.

[50] M. Lim, "Many clicks but little sticks: Social media activism in Indonesia", *Journal of Contemporary Asia*, vol. 43, no. 4, pp. 636–657, 2013. DOI: 10.1080/00472336.2013.769386.

[51] V. Pandey, S. Gupta, and H. Kim, "Exploring the role of technology affordance and social capital in promoting citizen's political participation on social media", *Pacific Asia Journal of the Association for Information Systems*, vol. 13, no. 4, 2021. DOI: 10.17705/1pais.13401.

[52] A. Johns, "This will be the whatsapp election: Crypto-publics and digital citizenship in Malaysia's ge14 election", *First Monday*, vol. 25, no. 12, 2020. DOI: 10.5210/fm.v25i12.10381.

[53] M. L. Weiss, "What will become of reformasi? ethnicity and changing political norms in Malaysia", *Contemporary Southeast Asia: A Journal of International and Strategic Affairs*, 1999. DOI: 10.1355/cs21_3f.

[54] M. Lim, "Sweeping the unclean: Social media and the bersih electoral reform movement in Malaysia", *Global Media Journal*, vol. 14, no. 27, p. 1, 2016.

[55] Y. H. Khoo, "Malaysia's 13th general elections and the rise of electoral reform movement", *Asian Politics & Policy*, vol. 8, no. 3, pp. 418–435, 2016. DOI: 10.1111/aspp.12273.

[56] J. Abbott, *Mahathir, Malaysia and the multimedia super corridor: Development catalyst, white elephant or cultural landmark?*, https://works.bepress.com/jason-abbott/29/, 2004.

[57] F. Loh and A. Netto, Eds., *Regime Change in Malaysia: GE14 and the End of UMNO-BM's 60-year Rule*. Strategic Information and Research Development Centre, 2018.

[58] J. Postill, "A critical history of internet activism and social protest in Malaysia, 1998-2011", *Asiascape*, vol. 1, no. 1–2, pp. 78–103, 2014. DOI: 10.1163/22142312-12340006.

[59] B. T. Khoo, "Networks in pursuit of a "two-coalition system" in Malaysia: Pakatan rakyat's mobilization of dissent between reformasi and the tsunami", *Southeast Asian Studies*, vol. 5, no. 1, pp. 73–91, 2016. DOI: 10.20495/seas.5.1_73.

[60] T. C. Chan, "Democratic breakthrough in Malaysia – political opportunities and the role of bersih", *Journal of Current Southeast Asian Affairs*, vol. 37, no. 3, pp. 109–137, 2018. DOI: 10.1177/186810341803700306.

[61] J. Hopkins, "Cybertroopers and tea parties: Government use of the internet in Malaysia", *Asian Journal of Communication*, vol. 24, no. 1, pp. 5–24, 2014. DOI: 10.1080/01292986.2013.851721.

[62] R. Tapsell, "The smartphone as the "weapon of the weak": Assessing the role of communication technologies in Malaysia's regime change", *Journal of Current Southeast Asian Affairs*, vol. 37, no. 3, pp. 9–29, 2018. DOI: 10.1177/186810341803700302.

[63] C. Leong, *Malaysia's bersih movement shows social media can mobilise the masses*, August 22, 2016.

[64] J. Funston, "Malaysia's 14th general election (ge14) -the contest for the Malay electorate", *Journal of Current Southeast Asian Affairs*, vol. 37, no. 3, pp. 57–83, 2018. DOI: 10.1177/186810341803700304.

[65] B. L. Liebman, "Watchdog or demagogue? the media in the Chinese legal system", *Columbia Law Review*, vol. 105, no. 1, pp. 1–157, 2005. DOI: 10.7916/d8j67gk4.

[66] S. Gan, "Virtual democracy in Malaysia", *Nieman Harvard, International Journalism*, vol. 56, no. 2, pp. 65–67, 2002.

[67] R. Tapsell, "The media freedom movement in Malaysia and the electoral authoritarian regime", *Journal of Contemporary Asia*, vol. 43, no. 4, pp. 613–635, 2013. DOI: 10.1080/00472336.2013.765138.

[68] L. T. Ghee, *Has the election already been stolen?*, April 20, 2018.

[69] J. Abbott, A. S. MacDonald, and J. Givens, "New social media and (electronic) democratization in east and southeast Asia: Malaysia and china compared", *Taiwan Journal of Democracy*, vol. 9, no. 2, pp. 105–137, 2013.

[70] J. Abbott, "Electoral authoritarianism and the print media in Malaysia: Measuring political bias and analyzing its cause", *Asian Affairs: An American Review*, vol. 38, no. 1, pp. 1–38, 2011. DOI: 10.1080/00927678.2010.520575.

[71] J. Funston, "The Malay electorate in 2004: Reversing the 1999 result?", in *ISEAS Publishing eBooks*, S. H. Saw and

K. Kesavapany, Eds., ISEAS Publishing, 2005, pp. 132–156. DOI: 10.1355/9789812305541-009.

[72] M. Trowell, *Sodomy II: The Trial of Anwar Ibrahim*. Marshall Cavendish International (Asia) Pte Limited, 2012.

[73] M. Trowell, *The Prosecution of Anwar Ibrahim: The Final Play*. Marshall Cavendish International (Asia) Pte Limited, 2015.

[74] J. Kurlantzick, *China's influence tactics in Malaysia—failure now, failure forever?*, March 3, 2023.

[75] H. Auto, *China to build, finance new east coast rail link in Malaysia*, October 31, 2016.

[76] S.-L. Wong, *Najib asks west to stop "lecturing" as Malaysia embraces china*, November 2, 2016.

[77] Today, *Johor sultan slams dr mahathir for playing 'politics of fear and race.'* January 17, 2017.

[78] Online, Star, *Kajd and powerchina sign rm30bil agreement for melaka gateway project*, November 29, 2019.

[79] Asia Maritime Transparency Initiative, *China's new spratly island defenses | Asia maritime transparency initiative*, December 13, 2016.

[80] Guardian Staff Reporter, *Malaysian PM Najib key figure in 1MDB corruption scandal, alleges cabinet minister*, July 14, 2017.

[81] ABC News, *Ethnic Malays openly denounce Chinese in rally organised by Malaysia's ruling party UMNO*, September 16, 2015.

[82] Y. S. Tan and H.-L. Teoh, "The development of Chinese education in Malaysia, 1952–1975: Political collaboration between the Malaysian Chinese association and the Chinese educationists", *History of Education*, vol. 44, no. 1, pp. 83–100, 2015. DOI: 10.1080/0046760x.2014.959073.

Aligning Business and Software Processes: GQM+Strategies Revisited

Luigi Lavazza, Sandro Morasca, Davide Tosi
Dipartimento di Scienze Teoriche e Applicate
Università degli Studi dell'Insubria
Varese, Italy
e-mail: {luigi.lavazza | sandro.morasca | davide.tosi}@uninsubria.it

***Abstract*—GQM has proven itself useful in supporting the definition and execution of software measurement plans. However, GQM has not been as effective in supporting the need for linking software measurement goals to higher-level goals, which typically originate in the business world. Hence, the GQM+Strategies technique was proposed to explicitly linking software measurement goals to the business world. However, little attention was given to the description of the business world. In this paper, we propose a method to precisely describe the business domain and its characteristics, the business goals, the strategies, their relationships with the software activities carried out to support the strategies, and how strategies are selected. We propose a way to firstly describe the business world, including business and software processes, and secondly to specify the measurements to be carried out. The proposed approach has been applied to a case proposed in the literature. The proposed approach proved very effective in supporting the investigation and descriptions of the business world. The creation of measurement plans according to the GQM+Strategies technique was greatly eased.**

Keywords-software development process; software process measurement; GQM; domain representation.

I. Introduction

In business contexts, it is of great importance that measurement practices are linked to high-level business goals in a clear and well reasoned way. A technique supporting such objectives was presented at SOFTENG 2025 [1]: it was proposed to use the GQM (Goal/Question/Metrics) technique [2]–[4]. That proposal is here extended, also with the help of a case study.

GQM has been successfully used for supporting the definition and execution of software measurement plans in a variety of industrial settings [5]–[9]. However, GQM has proven weaker in supporting the need for linking software measurement goals to higher-level goals of the organization for which the software is developed [10], [11]. Making the connection of software measurement to higher-level business goals explicit is very important for two reasons: justifying software measurement efforts and cost, and defining or using measures that effectively contribute to higher-level business-oriented decisions.

GQM+Strategies [10], [12]–[14] provides mechanisms for explicitly linking software measurement goals to business goals at different levels, up to the level of the entire business.

The scenario addressed by both GQM+strategies and this paper encompasses three elements: the business world, the measurement world, and the connections between the business and measurement worlds.

The business world (BW) is where the business operates, and includes the piece of the real world that is relevant for the business (including the market, users, stakeholders, competitors, etc.). In the BW, business goals are conceived, and strategies to achieve such goals are deployed. In most cases, strategies involve the usage of software, which very often needs to be specially developed to support a strategy. Strategies are hierarchical in nature, since implementing a strategy usually involves achieving a lower level goal, which, in turn, could require a strategy.

The measurement world (MW) is where measurement plans are specified, measures are defined, indicators (e.g., KPI) are computed. The measurement world is much more controllable than the business world. Accordingly, techniques and tools –like the GQM and related tools and methodologies– have been defined to support the work to be carried out in the measurement world.

The connections between the BW and MW represent the fact that the objects of measurement are in the business world and the data that support the evaluations performed in the measurement world are provided by the business world. Moreover, people from the two worlds need to agree on the measure definitions, how measurement is carried out, the meaning and expressiveness of indicators, etc.

GQM+Strategies highlights the relations existing between business goals and software development (or acquisition) within the BW and supports identifying and documenting the relationships between goals in the BW and measurement plans in the MW.

However, in GQM+Strategies, little attention is given to the description of the BW, with particular reference to the business goals and the strategies of interest [15]. In this respect, GQM+Strategies seems to inherit the weakness of GQM, which did not provide guidelines for modeling the relevant aspects of the software product and process that are the objects of measurement.

A clear understanding of the business domain, the rules and constraints that affect the business, the final goals of the stakeholders, and the cause-effect relationships that govern the business is of fundamental importance to devise effective strategies. Those who need to support such strategies by means of software and, then, measure the effectiveness of the software solutions and the implemented strategies, need to have access to explicit and clear descriptions of the BW. That is, they need to distinguish between what is given and cannot be changed, what is currently not true and must be achieved (the business goal), and what is the set of actions (the strategy) that have been planned to achieve the business goal.

In this paper, we propose a method to precisely describe the

BW, in terms of the business domain and its characteristics, the business goals, the strategies and their relationships with the software activities carried out to support the strategies. Then, i.e., after building an integrated and harmonic view encompassing both the business and software processes, we address the definition of business plans, using the GQM+Strategies proposal.

In other words, we believe that linking software development and business strategy is a rather complex task, so that identifying and modeling such link directly when defining measures is difficult and error-prone. Hence, we propose to first provide an integrated description of business and software processes (using the concepts of GQM+Strategies) and then to specify—via GQM plans—the measurements to be carried out. Quite noticeably, the latter activity becomes easier since the concepts of GQM+Strategies are used to describe the BW, so that identifying the business goals, the context, the strategies, etc. is straightforward.

Note that in this paper we do not propose any brand new methods for specifying business needs and software requirements. Instead, we build on existing proposals. Specifically, in addition to GQM+Strategies, we borrow ideas from Jackson's work [16]–[19] on requirements and domain representation. Thus, we are able to propose an approach that is simple and fairly easy to understand, yet powerful and applicable in practice.

The remainder of the paper is organized as follows: Section II summarizes the GQM+Strategies method and highlights the need for better models of the BW; an example originally proposed by Basili et al. [13] is also introduced. Section III proposes a (meta)model to represent the hierarchy of requirements in the BW. Section IV discusses the selection of strategies to achieve goals. Section V shows how the proposed approach can be applied to describe the BW for the example described in the Section II. Section VI discusses what to measure and describes how to links goals, strategies and the knowledge of the business world to GQM measurement plan . Related work is commented in Section VII. We conclude and we draw some directions of further investigation in Section VIII.

II. GQM+Strategies

Here, we first summarize the method as proposed by Basili et al. [10], [12], [13]. Then, we comment on the need for descriptions of the BW that are more precise and systematic than those proposed by the GQM+Strategies method.

A. A concise introduction to GQM+Strategies

GQM+Strategies aims to address the weaknesses of existing goal-oriented approaches by providing explicit links among organizational levels in a flexible manner, to tailor the approach to the organization's specific needs and objectives [12]. The proposed conceptual components are illustrated in Figure 1, taken from [12]:

- Business goals are specific organization's goals that call for software development.
- Context factors account for environmental conditions that affect both the goals and how they are pursued.
- Assumptions concern the parts of the context that are not known with certainty.
- Strategies indicate how a goal is pursued. The implementation of a strategy may involve achieving lower-level goals.
- Interpretation models indicate how to interpret data to determine if the goals at all levels have been achieved.
- A Goal+Strategies element groups the business goal, the associated strategy, and the context information and assumptions at a given level.

A single GQM goal measures a Goal+Strategies element. An example of GQM+strategies model is given in Figure 2, taken from [13]. The considered company operates in a market that is becoming highly competitive, so that there is a need to safeguard the company's place in the market, i.e., to keep the current customers. To this end, generating customer loyalty is necessary. This can be achieved by improving customer satisfaction with the next product, so business goal "increase customer satisfaction by 10%" is defined.

An analysis revealed that many customer complaints are due to product reliability problems. After considering several possible strategies, it was decided that the most promising way to increase customers' satisfaction is to "test reliability in".

In order to test reliability in, the software test processes are examined and potential lower-level goals are identified. The company has discovered a new system test process that seems appropriate for their context and that can decrease the total number of customer complaints by 10% by reducing customer-reported software field defects (i.e., those that slip by system test) by 20%. Thus, the second-level goal is to improve system test effectiveness by 20%. Because there is a new suitable system test process, the one and only strategy is to introduce the new system test process.

Based on historical defect slippage data, the company assumes that reducing slippage by 20% reduces reported defects by 20%. So, the lower level goal is to apply the new system test method in order to see if it reduces defect slippage by at least 20% and generates the necessary improvement to customer complaints.

B. On the need for better models of the Business World

The need for clarifying the BW anticipated in the Introduction can be illustrated by means of the example given in [13] and reported in Figure 2.

First, the boundaries of the BW model should be explicitly defined. Similarly, it should be clarified why some elements of the BW are in the model, while others have been excluded. In fact, given a business goal, it is always possible to wonder from where it originates, what business needs led to the definition of such goal, etc. At the opposite end, a goal that is at the ground level in a model can always call for a strategy. In fact, any goal that can be pursued in two or more different ways can be associated with a "strategy" that simply indicates which of the several possible implementation ways has been chosen.

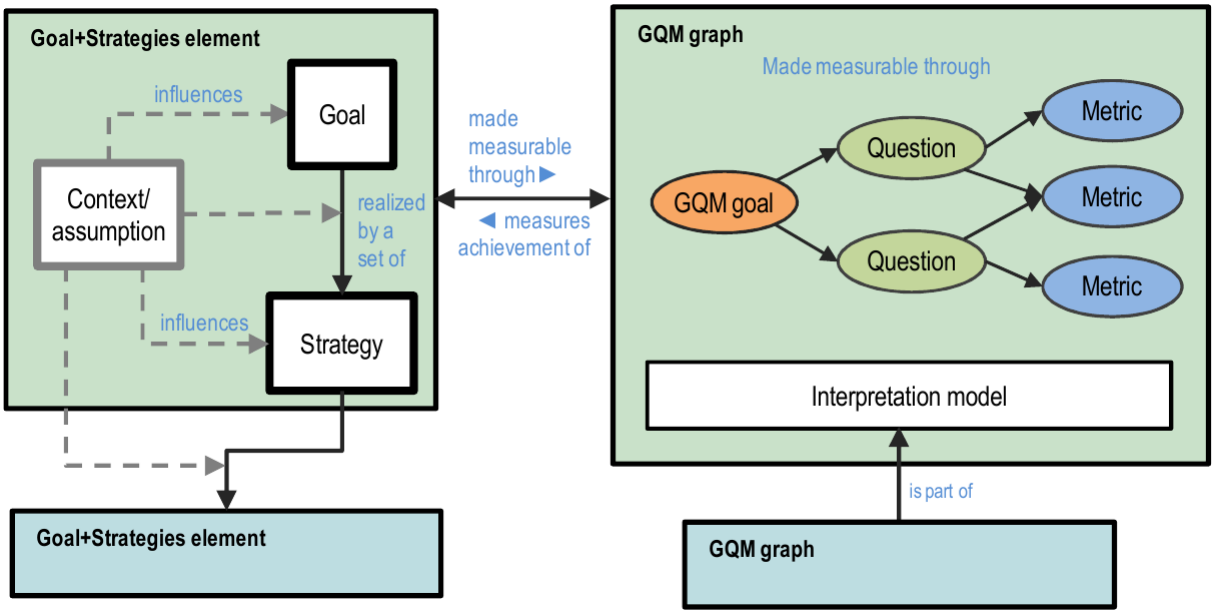


Figure 1. GQM+Strategies components [12].

Concerning the top-level goal (increasing customer satisfaction by 10%), in Figure 2, the presence of context C1 and assumption A1 suggests that such context information and the related assumptions were used to formulate the goal. However, the specific problem to which C1 and A1 were applied is not mentioned, so we do not know if there is an even higher-level goal that can be reached by pursuing the top-level goal appearing in in Figure 2. In fact, one might infer that the (unknown) higher-level business goal is increasing customers' loyalty, or just preserving the current market share, since in a competitive market, improving customers' satisfaction could be necessary not to reduce a company's market share. C1 and A1, if applied to such goal, would result in the decision of increasing customer satisfaction by 10%. In turn, this higher-level business goal may come from an even higher-level business goal (e.g., ensure the company's long-term viability). Several levels of higher-level goals may be possible.

To stop this upward chain of goals, the top-level goal should be given as an "axiom," and no context or assumptions should be provided to justify it. Otherwise, one could wonder for what specific purpose context or assumptions are applied, thus looking for a further upper level.

On the contrary, at the bottom level, the basic strategy should be either sufficiently simple to require no further refinement and it should be measurable, as any strategy in the GQM+Strategies approach.

Another fundamental observation is that several different strategies can possibly satisfy a given business goal. For instance, customers' satisfaction can be increased in several different manners: increasing the reliability of products is surely a way, but it could be possible to decrease prices, to add functions, to improve efficiency, etc. The criteria for choosing a strategy over others are not given in Figure 2. This is a rather severe limit of the BW modeling in GQM+Strategies. In some cases, strategies could be constrained by the context (e.g., decreasing the price of the product could just be impossible), but strategies more often derive from the preferences and the knowledge (obtained via market analysis and the like) of the people in charge of decisions. In such cases, explicitly recording the decision criteria that lead to selecting a strategy would be beneficial, since decision criteria could play a very important role in the evaluation of strategies.

In the considered case, suppose that two possible strategies were viable: a) give discounts to customers; b) increase customers' support. Suppose also that option b) was chosen because it was considered less expensive for the company. This decision can be carried out based for instance on market data and simulations (a sort of "what if" analysis) for both strategies. Thus, if all this were documented, we could evaluate if the strategy selection criterion is sound, according to the knowledge and models available before choosing strategy b). After the selected strategy is executed, we could verify whether the criterion worked as expected, i.e., if customer loyalty increased when increasing customer support. In any case, since we have already evaluated what would have happened if we had chosen strategy a), according to available data, we can compare the actual results obtained with strategy b) and the predicted results with strategy a). In any case, the criterion "a) costs more than b)" would be explicit and verifiable. Over time, by recording the decisions made, their rationales, and the results obtained, we can reach a reliable evaluation of the strategy selection criterion that can be recorded (e.g., in the Experience Factory [20]) as an asset for the organization for future use.

A further observation is that the connections between strategies and the corresponding sub-goals in Figure 2 are not all well defined. For instance, reducing defect slippage must

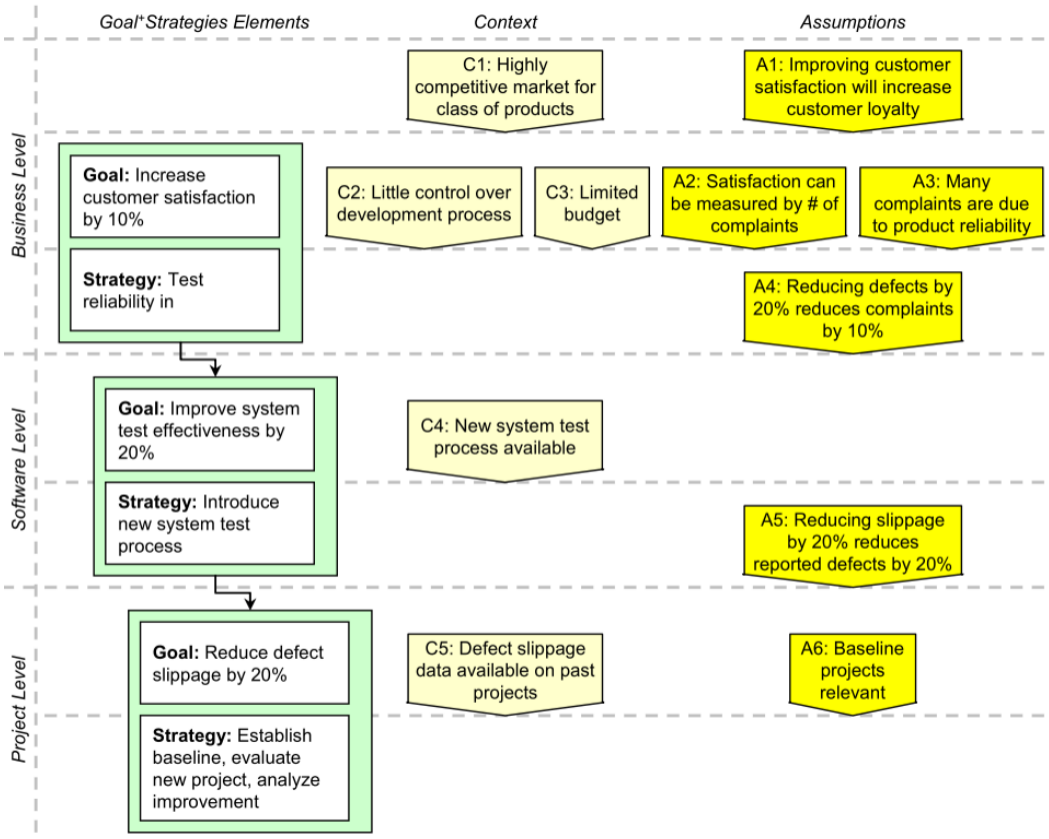


Figure 2. Goals, strategies, context factors and assumptions [13].

be a property of the new system test process addressed by the software-level strategy. In fact, "introducing a new system test process" is not a reasonable strategy by itself, since the new process could be worse than the existing one. An effective strategy consists in "introducing a new system test process that reduces defect slippage by 20%." The origin of the confusion here is probably that in this case the project-level goal could be identical to the software-level strategy. In other words, to implement a strategy that requires a new system test process that reduces defect slippage by 20% what is needed is just a project whose goal is to introduce a new system test process that reduces defect slippage by 20%. The fact that a strategy at a given level becomes a goal as-is at the immediately lower level should not be surprising. Finally, it should be noted that implementing a given strategy could require multiple sub-goals. For instance, a Strategy could require achieving sub-goals both on the technical side (e.g., introducing tools, platforms, etc.) and on the competence side (e.g., acquire some skill or knowledge). This situation—not shown in Figure 2—is quite common in practice.

III. Requirements Hierarchies

Some work has been done in the area of representing the hierarchies of requirements ranging from high-level business requirements to software development requirements. Some of the existing proposals also address the precise (possibly formal) description of the requirements and the domains in which they exist.

We here propose a (meta)model for representing in a quite systematic and rigorous way the hierarchy of requirements in the BW and (in the next section) their links to the measurement plans.

A. Describing requirements: Jackson's method

In this paper, we use the concepts for requirements specification proposed by Jackson [16]–[19].

Jackson noted that the relationship between user requirements and the specifications of a hardware/software machine that has to satisfy such requirements in a given environment can be described as follows [21]:

$$E, S \vdash R \tag{1}$$

In (1), R indicates the stakeholder's requirements, i.e., "conditions over the phenomena of the environment that we wish to make true by installing the machine" while E expresses "conditions over the phenomena of the environment that we know to be true irrespective of the properties and behaviour of the machine" [17]. S is the specification of the machine, expressed in terms of the phenomena that are shared by the environment and the machine. In practice, S is specified in terms of I/O elements: in Jackson's terminology, phenomena controlled by the environment and visible to the machine (i.e.,

inputs) and phenomena controlled by the machine and visible to the environment (i.e., outputs).

Formula (1) states that if the environment in which the machine is located behaves as specified in E and the specifications S are satisfied by the machine and the environment, then requirements R are satisfied. The logical entailment $A \vdash B$ states that from assuming A we can prove B; hence, entailment is often called provability. It is important to note that the level of formality of formula (1) depends on the formality of the descriptions E, S and R. If E, R and S are described formally it is possible to prove that the truth of R descends from the truth of E and S, while informal descriptions allow only for argumentations, which are deemed sufficient in most case, though.

B. Describing requirements hierarchies

Given a context and a goal, the strategy is the "solution" that—in the given context and under the given assumptions—satisfies the goal.

Using Jackson's concepts and notation, the statement above can be written as follows.

$$Context, Strategy \vdash Goal$$

where Context is the description of the business domain, including all the knowledge that is relevant with respect to the goals currently considered, Goal is the description of what is desired by the business actors, and Strategy is the solution that has been devised to achieve the goal.

In Jackson's terminology, the context is given, thus it is "indicative." More precisely, in GQM+Strategies [12], the context includes:

- Factors (known with certainty);
- Assumptions (uncertain).

Accordingly, we could say that

$$Context_F, Context_A, Strategy \vdash Goal$$

where $Context_F$ is the description of the factors known with certainty and $Context_A$ is the description of the assumptions. It could be observed that also part of the context can be controlled or changed: for instance, the employees working in the considered environment can be instructed to behave in a given way, devices can be installed, etc.. Thus, this part of the context is not really indicative, since changing it could actually be part of a strategy.

The Goal is "optative," i.e., it represents something that is not currently true, but needs to be made true. In fact, the application of the Strategy in the Context is the means by which the Goal is satisfied.

The Strategy is clearly optative, since in general the Goal can be achieved via several different strategies. More precisely, the initial situation can generally be represented as follows:

$$Context, ? \vdash Goal$$

In this formula, the question mark explicitly indicates that in a given (known) context there is a goal (i.e., some desirable conditions currently not holding), but how to achieve it is yet to be defined.

Once the Strategy has been described, i.e., we have decided what has to be achieved, it is necessary to specify how it should be achieved. This is why goals and strategies form hierarchies (as in Figure 1): implementing a strategy in general requires the achievement of some lower-level goal. For this purpose, a lower-level strategy is required, which could require the achievement of an even lower-level goal, etc. This type of requirements hierarchies is described in [22], using Jackson's notation.

$$Context, Strategy \vdash BusinessGoal$$
$$Context, LowerLevelGoal \vdash Strategy$$
$$Context, LowerLevelStrategy \vdash LowerLevelGoal$$

The LowerLevelGoal specifies what we can do to realize the Strategy: reaching LowerLevelGoal in the Context is a sufficient condition for the realization of the Strategy. However, LowerLevelGoal is itself a goal, so it is also necessary to specify how the LowerLevelGoal should be achieved. To this end, we need to devise a LowerLevelStrategy to reach LowerLevelGoal as shown in the last logical entailment above.

IV. Selecting a Strategy

Different strategies are characterized by different costs, effectiveness, risks, and benefits, so that choosing a strategy (i.e., exercising the option) implies that multiple characteristics of multiple strategies may need to be assessed. Therefore, in addition to the Goal, a Figure of Merit (FM) exists, whose value depends on the Context and the Strategy. The FM can be used in two ways. First, a constraint can be set on the FM. For instance, if cost is the FM, we can consider acceptable only strategies whose cost is below a specified cost threshold. Second, the FM can be used to comparatively assess different strategies, based on a Preference Criterion (PC) that ranks alternatives based on their corresponding values of FM. The PC may be a straightforward one when the FM is a single-objective one. However, FMs are often multiple-objective: for instance, a double-objective FM may address effort and development time. The application of the PC results in general in a partially ordered set of strategies, as some strategies may be deemed equivalent as for their FMs.

Making the FM and PC explicit shows that the selection of a strategy is not based only on the Goal; instead, it involves the optimization of characteristics that do not necessarily appear in the Goal. For instance, take the business Goal in the example, which should be interpreted as "Increase customer satisfaction by at least 20%." This Goal sets a constraint on the set of possible strategies used to reach it, but by no means does it explicitly indicate how to choose among competing strategies that satisfy it. In principle, one could choose any Strategy that satisfies the Business Level Goal in the given Context, regardless of the cost. However, in practice, the Strategy that minimizes the cost is likely to be preferred over the others.

Also, making the FM and the PC explicit provides guidance in the building of effective strategies, when no previously used strategies are available, or in the tailoring of existing ones or when there is a significant level of uncertainty, which is always present when making decisions. If so, we may not be able to identify the optimal Strategy with certainty, but the FM and the PC will help us at least reduce the set of strategies.

Summarizing, the FM and the PC need to be made explicit so that all ambiguities are removed as to why a specific Strategy is selected. Also, the analysis of the results obtained in the field will allow us to refine our decision processes.

V. Describing the Business World

To illustrate the notation described in Section III, in this section, we describe the example illustrated in Figure 1 using the proposed notation. In describing the business world and how to cope with the given high-level goal, we also show how to solve the problems with the GQM+Strategies descriptive power discussed in Section II.

The high-level objective is *"Preserve market share."* This objective is given and is not under discussion. It serves as an axiom for the following reasoning.

Having the goal, we have to decide how to pursue it, i.e., we have to define a strategy that (hopefully) will let us achieve the goal. To define the strategy, we consider the following knowledge of the business world:

High competitive markets, User satisfaction not high $\vdash$
Loss of market share

Since the market competitiveness is out of our control, to avoid losing market share we have to achieve high customer satisfaction. In other words, we adopt a strategy consisting in improving user satisfaction. Note that loss of market shares could be caused also by other factors than user satisfaction, but these other factors are not considered in this example.

So, *Improving user satisfaction* is a new goal. Again, we have to devise a strategy to achieve this goal. To such end, we can exploit the knowledge that satisfaction depends on reliability (the more reliable is a product, the more satisfied is the customer), as stated by the following entailment:

Increased product reliability $\vdash$ *Improved customer satisfaction*

So, increasing product reliability (i.e., *Reducing defects* in the released product) is our new goal. To pursue this objective, we exploit the knowledge that effective tests decrease defects.

Effective testing $\vdash$ *Low defect rate*

So, the new objective is *Increase test effectiveness*. To pursue this objective, we exploit the knowledge that

New system test process available,
New system test process usage $\vdash$ *Increased test effectiveness*

Therefore, adopting the new system test process is the chosen strategy. We can suppose that the adoption of the new system test process is a "ground activity" that does not need to be further discussed. Otherwise, we could proceed as above, by considering what is needed to introduce the new system test approach in the current development process, etc.

Note that the description above is coherent with the GQM+Strategy definition. This is evident from Figure 3, which describes a fragment of the requirements hierarchy as a GQM+Strategy element: its can be observed that the Goal, Strategy and Context/assumptions are described via the proposed notation borrowed from Jackson.

Now, a quite important point with the proposed notation is that Jackson does not prescribe how to specify individual statements. For instance, *"Improving user satisfaction"* could be specified in plain English or by means of some type of temporal logic. Quite interestingly, the same happens with GQM, where the processes or products that are object of measurement can be described in any possible way.

We maintain the approach of Jackson and GQM, and do not suggest any specific notation: each practitioner is free to adopt the notation he/she is most familiar with.

Other important observations are the following.

As already mentioned, given a strategy, we have to specify how the strategy is carried out. That is, we need to define a lower level goal that represents the implementation of the strategy. While the GQM+strategies method is not very clear on this point, with our approach the procedure is quite straightforward: the relationship between Strategy and sub-goals is the same as between Goal and Strategy.

It can be noticed that in formula $Context, Strategy \vdash BusinessGoal$ the Strategy is optative, i.e., we are free to choose one among the possibly multiple strategies that support the achievement of the goal. On the contrary, in formula $Context, LowerLevelGoal \vdash Strategy$, the strategy has become indicative, while the lower level goal is optative, since there are potentially multiple ways to implement the strategy: for instance, *Improved customer satisfaction* could be the consequence of other factors than *Increased product reliability*. These observations are coherent with the fact that proceeding from the business goal level to the operational goals at the lowest level involves a sequence of decisions. The description method we propose is suitable to represent the progress of the decisional process as well as the cause-effect relationships that link goals and strategies at the different levels.

Another important observation is that formulae like

Effective testing $\vdash$ *Low defect rate*

can be written at different levels of detail. For instance, we could distinguish different types of defects. Also in this case, the proposed method can be used at the level of detail considered suitable by the user: e.g., one could specify what "effective test" means and what are the expected effects on the different types of defects.

A final consideration is that Basili et al. suggest three levels, namely Business, Software and Project [12], while we do not limit the number and type of levels.

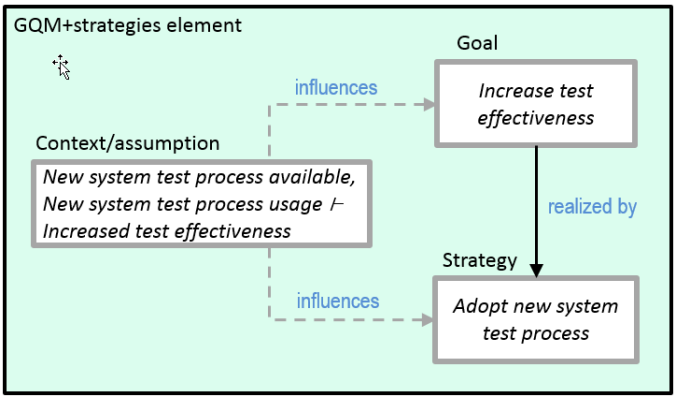


Figure 3. A GQM+Strategy element from the case study.

VI. What Should Be Measured (and How)

Basili et al. provide the following indications for measurement [12]:

> *Associated with each GQM+Strategies element is a measurement plan that uses the GQM measurement and evaluation framework to specify how to evaluate the goal, what data to collect, and how to interpret that data. The nodes of each GQM graph consist of a measurement goal, which describes what knowledge needs to be gained from the measurement activity; a set of questions to be answered; the metrics and data items required to answer the questions; and an interpretation model that specifies how the data items are to be combined and what the criteria are for determining the goal's success.*

With respect to the situation described in Figure 1, formula

$$Context, Strategy \vdash Goal$$

provides clearer indications about what should be measured. In fact, while in [12] a single GQM plan is connected to a Goal+Strategy element, our method naturally suggests specific measurement plans for each part of the entailment:

- Context: if the context description contains assumptions, it is generally a good practice to measure to what extent the assumptions are true.
- Goal: of course, we want to know to what extent the goal has been achieved. To this end, a GQM plan is typically attached to the business goal.
- Strategy: we want to know to what extent and how well the strategy has been applied. So, a specific GQM plan is typically defined for the strategy.

For sure, we want to measure the Figure of Merit associated to a given entailment. In some cases, we could even have multiple Figures of Merit, each one representing a specific point of view. For instance, we could have a Figure of Merit for top management and another one for the project manager. Measuring the Figure(s) of Merit usually requires measuring the elements (the Context, the Strategy and possibly the Goal) to which a Figure of Merit refers. However, it must be noted that very often a Figure of Merit concerns properties (e.g., the amount of resources used to implement a Strategy, or the time taken to complete the activities involved in a given Strategy) that belong to a sort of meta-level, and are possibly not considered in the "basic" measurement of Strategy. The quantification of Strategy selection criteria usually does not call for additional measures, instead it is just a function of the computed Figure(s) of Merit.

As an example, let us consider the entailment *New system test process usage* $\vdash$ *Increased test effectiveness* from Section III: evaluating the Figure of Merit involves measuring properties like the cost of testing, the increase of competence needed to conceive tests, the cost and the learning curve of tools and practices used for testing, etc.

The entailment is usually assumed to be true. In other words, it is believed that the devised Strategy, correctly applied in the given Context, causes the full achievement of the Goal. However, it may happen that the Goal does not follow from the Context and Strategy: measuring also this fact is therefore advisable. This usually involves verifying the connections between properties of the processes and products addressed by the Strategy and processes and products considered in the Goal. For instance, one of the conditions that make the entailment *Effective testing* $\vdash$ *Low defect rate* true is that the results of testing are fed to an effective debugging activity.

The interpretation model mentioned in [12] is clearly of great importance, since the whole interpretation of the collected data depends on it. Nevertheless, in [12] it is not specified how the interpretation model should be defined, it being delegated as part of the GQM plan. This is not advisable, in that the GQM itself is generally more oriented to refining goals into metrics than in prescribing how the collected data would be interpreted.

With our approach, the interpretations are generally made apparent by the formulae. Moreover, we do not have multiple GQM plans and graphs, as in Figure 1; instead we have a single plan, with clearly interconnected elements, as shown in Figure 4 (which schematically represents a portion of the requirements hierarchy).

The connections between a strategy and its lower level goals

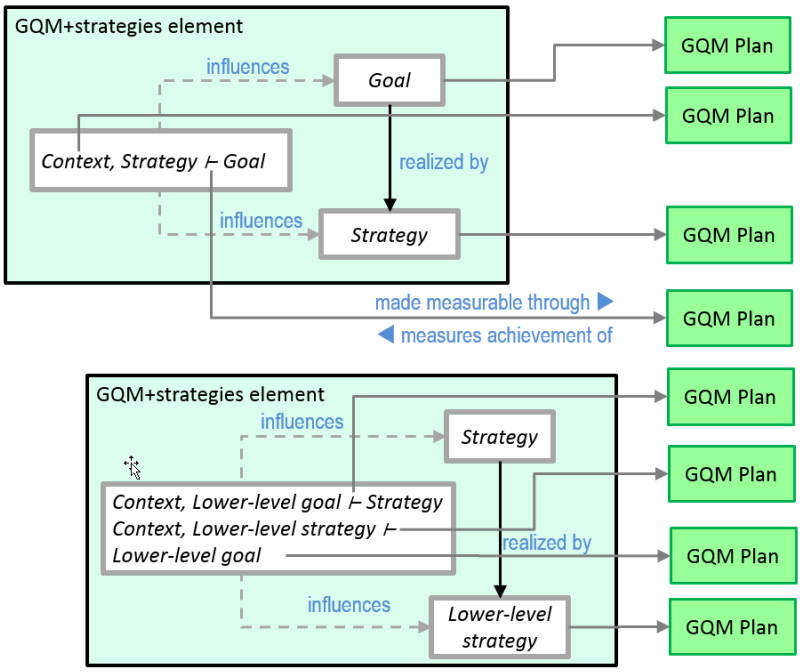


Figure 4. Our proposal for measurement.

are not emphasized by Basili et al., while they are clearly represented and measured in our approach.

Figure 5 illustrates the connection of GQM+Strategy elements to GQM plans with our approach: it can be observed that every component of the considered GQM+Strategy element, i.e, every goal, strategy, entailment, etc., is connected to one or more GQM plans. In this way, the definition of GQM plans is straightforward.

VII. Related Work

The weakness of GQM in describing the software product or process that is the object of measurement were overcome by coupling GQM-compliant measurement tools with tools for modeling the product and process [23]. The work described here can be seen as a logical prosecution of that work, in that here we provide the basis for coupling reasoning on business goals, user requirements, software development and –finally– measurement.

The need for linking business processes and Goal/Question/Metric paradigms has been felt since 2004 [24]. In [24] the authors define a measurement framework to support process analysts in assessing business processes by means of the GQM paradigm, to find useful indications about process performance, critical elements, change impact, and expected improvement. In our approach, the focus moves from a way to assess the quantitative and qualitative aspects of a business process to a way to precisely (possibly formally) describe business processes in a manner that is compliant with the GQM paradigm. The precise description of the business world and of company goals eases both the measurement of process aspects and the evaluation—both quantitative and qualitative—of the business and technical aspects of the process.

GQM+Strategies has been introduced for the first time in [10] to extend the GQM approach with the capability to create measurement programs that ensure a link between business goals and strategies, software goals, and measurement goals. The approach has been supported by the SAS tool to improve the definition of the context, assumption, and strategies [25]. In our paper, we adopt the extensions proposed in [10] to go further in the direction of representing the BW processes that are to be connected with GQM+Strategies. Our approach makes the representation of relevant relationships explicit, independently from the GQM.

In [26], the GQM+Strategies approach is adopted to perform business value analysis and to identify success/critical business goals. The paper clearly states that the various aspects of business value expressed and defined by goals require the knowledge and experience of the stakeholders to identify what elements (context, assumptions, strategies, goals) are valuable and appropriate for the company's success. In our paper, we aim at improving the process of describing the BW, in terms of the business domain, characteristics, goals, strategies and relationships with the software activities.

In [15], the author notes that the business level should be mapped into a Conceptual/Strategic level to clearly define the scope of the Business level in a generic way (i.e., outside the boundary of the software domain): the conceptual level is actually the highest organizational abstraction where an organization determines how to succeed in those activities that are strategic for the existence of the organization itself. This kind of mapping is quite easy with our approach.

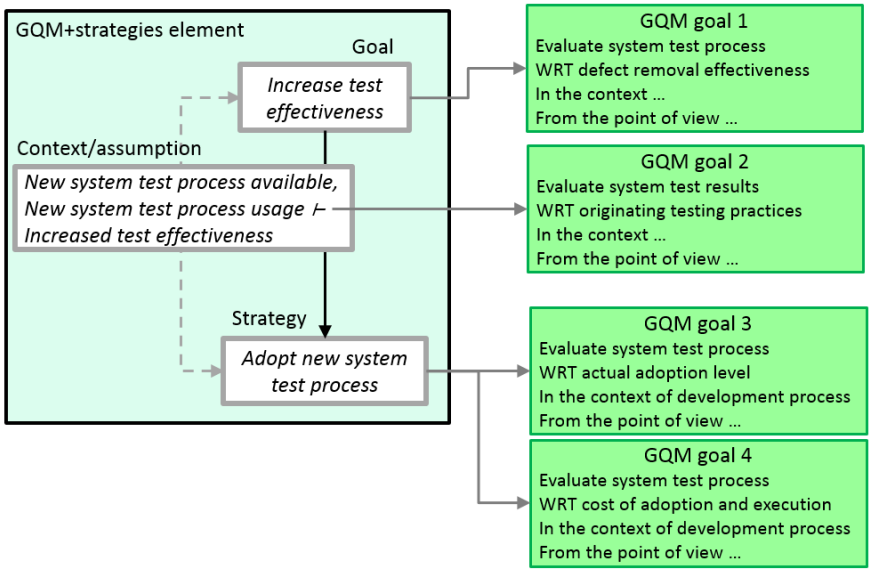


Figure 5. Linking GQM plans to GQM+strategy element components.

VIII. Conclusions and Future Work

In this paper, we propose a method to better link the GQM+Strategies methodology with the business goals of an organization. Our proposal is based on using Jackson's ideas on domain representation, and allows for the precise description of the business domain, the business goals, the strategies, and their relationships with the software activities carried out supporting the strategies, and how strategies are selected.

Thus, the proposal of this paper does not consist in inventing a new technique or notation, but in using two existing techniques to make their joint use applicable in practice. The proposal also makes it possible to clearly and explicitly describe and therefore record the rationale behind the selection of strategies. A Figure of Merit (in addition to a Goal) of practical interest needs to exist for the evaluation of strategies in a given Context. A Preference Criterion must be defined so the different strategies can be ranked according to the values of their Figure of Merit. The approach proposed in this paper has been demonstrated by using it on a case proposed in the literature on GQM+strategies.

The proposed technique is meant to describe buiness environments and strategies, in relation with software activities and software artifacts. As such, it can be used in multiple context and for multiple objectives. For instance, the proposed tecnique is suitable to support process improvement initiatives, irrespective of the adopted framework (e.g., CMMI, ISO standards, etc.)

A significant amount of future work remains do be done, including applying the approach to larger application cases and developing supporting tools to be integrated with existing GQM tools [23], [27]–[30].

Acknowledgment

The work reported here was partly supported by Fondo per la Ricerca di Ateneo, Università degli Studi dell'Insubria.

References

[1] L. Lavazza, S. Morasca, and D. Tosi, "Putting business goals in context for measurement", in *SOFTENG 2025 : The Eleventh International Conference on Advances and Trends in Software Engineering*, 2025, pp. 8–13.

[2] V. R. Basili and H. D. Rombach, "The TAME project: Towards improvement-oriented software environments", *IEEE Transactions on Software Engineering*, vol. 14, no. 6, pp. 758–773, 1988.

[3] V. R. Basili and D. M. Weiss, "A methodology for collecting valid software engineering data", *IEEE Transactions on Software Engineering*, no. 6, pp. 728–738, 1984.

[4] V. R. Basili, G. Caldiera, V. R. Basili, and H. D. Rombach, "Goal question metric paradigm", *Encyclopedia of Software Engineering*, vol. 1, no. 528-532, p. 6, 1994.

[5] A. Fuggetta *et al.*, "Applying GQM in an industrial software factory", *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 7, no. 4, pp. 411–448, 1998.

[6] A. Birk, R. Van Solingen, and J. Jarvinen, "Business impact, benefit, and cost of applying GQM in industry: an in-depth, long-term investigation at Schlumberger RPS", in *Proceedings Fifth International Software Metrics Symposium. Metrics (Cat. No. 98TB100262)*, IEEE, 1998, pp. 93–96.

[7] R. Van Solingen and E. Berghout, "Integrating goal-oriented measurement in industrial software engineering: industrial experiences with and additions to the Goal/Question/Metric method (GQM)", in *Proceedings Seventh International Software Metrics Symposium*, IEEE, 2001, pp. 246–258.

[8] L. Lavazza, "Multi-scope evaluation of public administration initiatives in process automation", in *The European Conference on Information Systems Management*, Academic Conferences International Limited, 2011, p. 294.

[9] L. Lavazza and M. Mauri, "Software process measurement in the real world: Dealing with operating constraints", in *Software Process Change: International Software Process Workshop and International Workshop on Software Process Simulation and Modeling, SPW/ProSim 2006, Shanghai, China, May 20-21, 2006. Proceedings*, Springer, 2006, pp. 80–87.

[10] V. Basili *et al.*, "GQM+strategies—aligning business strategies with software measurement", in *First International Symposium*

on Empirical Software Engineering and Measurement (ESEM 2007), IEEE, 2007, pp. 488–490.

[11] L. C. Briand, S. Morasca, and V. R. Basili, "An operational process for goal-driven definition of measures", *IEEE Transactions on Software Engineering*, vol. 28, no. 12, pp. 1106–1125, 2002.

[12] V. R. Basili *et al.*, "Linking software development and business strategy through measurement", *Computer*, vol. 43, no. 4, pp. 57–65, 2010.

[13] V. Basili *et al.*, "Determining the impact of business strategies using principles from goal-oriented measurement", in *Business Services: Konzepte, Technologien, Anwendungen. 9. Internationale Tagung Wirtschaftsinformatik*, Osterreichische Computer Gesellschaft, 2009.

[14] V. Mandić, L. Harjumaa, J. Markkula, and M. Oivo, "Early empirical assessment of the practical value of GQM+ Strategies", in *New Modeling Concepts for Today's Software Processes: International Conference on Software Process, ICSP 2010, Paderborn, Germany, July 8-9, 2010. Proceedings*, Springer, 2010, pp. 14–25.

[15] S. A. Sarcia', "Is GQM+Strategies really applicable as is to non-software development domains?", in *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, 2010, pp. 1–4.

[16] C. A. Gunter, E. L. Gunter, M. Jackson, and P. Zave, "A reference model for requirements and specifications", *IEEE Software*, vol. 17, no. 3, pp. 37–43, 2000.

[17] M. Jackson, "The meaning of requirements", *Annals of Software Engineering*, vol. 3, no. 1, pp. 5–21, 1997.

[18] M. Jackson, *Problem Frames: Analyzing and structuring software development problems*. Addison-Wesley Longman Publishing Co., Inc., 2000.

[19] M. Jackson and P. Zave, "Deriving specifications from requirements: An example", in *Proceedings of the 17th International Conference on Software Engineering*, 1995, pp. 15–24.

[20] G. Caldiera, V. R. Basili, and H. D. Rombach, "The experience factory", *Encyclopedia of Software Engineering*, vol. 1, pp. 469–476, 1994.

[21] M. Jackson, "Software specifications and requirements: A lexicon of practice, principles and prejudices", *Addison-Wesley, New York*, 1995.

[22] L. Lavazza, "Business goals, user needs, and requirements: A problem frame-based view", *Expert Systems*, vol. 30, no. 3, pp. 215–232, 2013.

[23] L. Lavazza and G. Barresi, "Automated support for process-aware definition and execution of measurement plans", in *Proceedings of the 27th International Conference on Software Engineering*, 2005, pp. 234–243.

[24] L. Aversano, T. Bodhuin, G. Canfora, and M. Tortorella, "A framework for measuring business processes based on GQM", in *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*, IEEE, 2004, 10–pp.

[25] V. Mandić and M. Oivo, "SAS: A tool for the GQM+strategies grid derivation process", in *Product-Focused Software Process Improvement: 11th International Conference, PROFES 2010, Limerick, Ireland, June 21-23, 2010. Proceedings 11*, Springer, 2010, pp. 291–305.

[26] V. Mandić, V. Basili, L. Harjumaa, M. Oivo, and J. Markkula, "Utilizing GQM+Strategies for business value analysis: An approach for evaluating business goals", in *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, 2010, pp. 1–10.

[27] C. Lofi, "Continuous GQM", Ph.D. dissertation, Master's thesis, Technical University Kaiserslautern, Germany, 2005.

[28] M. Hoffmann, A. Birk, F. van Els, and R. Kempkens, *GQM aspect v1.0 User manual*, 1996.

[29] J. C. Abib and T. G. Kirner, "A GQM-based tool to support the development of software quality measurement plans", *ACM SIGSOFT Software Engineering Notes*, vol. 24, no. 4, pp. 75–80, 1999.

[30] P. Parviainen, J. Jarvinen, and T. Sandelin, "Practical experiences of tool support in a GQM-based measurement programme", *Software Quality Journal*, vol. 6, pp. 283–294, 1997.

Radar Area Chart as a Framework for Quantitative Analysis of Electronic Noses in Monitoring Water Stress in Plants

Paulo S. de P. Herrmann[1], Matheus Santos Luccas[2]
[1]Embrapa Instrumentation (CNPDIA), São Carlos, SP, Brazil
[2] Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos, SP, Brazil
e-mail: paulo.herrmann@embrapa.br, matsluccas@gmail.com

***Abstract*—Water stress is a major factor limiting the productivity of soybean (Glycine max L.) worldwide. Early detection is crucial for implementing timely irrigation strategies. Electronic noses (E-noses) offer a promising, non-invasive approach for monitoring plant gas emissions. This work extends our previous publication and proposes the quantitative use of radar chart areas as a novel metric to evaluate and compare the sensitivity of individual E-nose sensors in detecting water stress. By calculating the polygon area formed by the normalized response peaks of a six-sensor array, we transform multivariate sensor data into a single, comparable index. This approach was applied to data from soybean plants subjected to controlled irrigation and water stress conditions over 21 days. The results demonstrate that the radar chart area metric effectively captures the temporal progression of stress, showing a distinct divergence between irrigated and non-irrigated plants after the onset of water stress. The proposed area-based metric provides a comprehensive, quantitative tool for sensor performance evaluation, enhancing the interpretation of E-nose data. This methodological advancement not only validates the radar chart area as a robust indicator of plant stress but also paves the way for more precise, data-driven decisions in precision agriculture applications.**

***Keywords*—Radar charts; Electronic nose; Precision agriculture; Water stress; Volatile Organic Compounds.**

I. Introduction

This work extends our previous publication [1] and builds upon recent advances that have introduced electronic noses (E-noses) as innovative tools for monitoring plant health. E-noses consist of sensor arrays capable of detecting specific patterns of Volatile Organic Compounds (VOCs) emitted by plants under stress conditions (see Figure 1). Changes in VOC profiles, which can serve as early indicators of water stress even before visual symptoms appear [2], are a key focus of this extension.

Soybean (*Glycine max* L.) is a vital crop, serving as a key source of protein and oil for human consumption and animal feed. Water stress remains one of the most critical abiotic stresses affecting soybean growth and yield, leading to global economic losses. Traditional methods for detecting water stress involve physiological measurements and visual assessments, which can be labor-intensive and subjective.

E-nose technology has been extensively evaluated in experiments employing statistical techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to reduce the dimensionality of the collected data [4]. Among these, the dynamic mode has been a distinguished method for identifying the most informative features that distinguish between varying stress levels. In this context, the peak response values of semiconductor sensors during VOC adsorption and desorption have been utilized as effective features for such analysis [2], with radar charts serving as a visualization tool for this purpose.

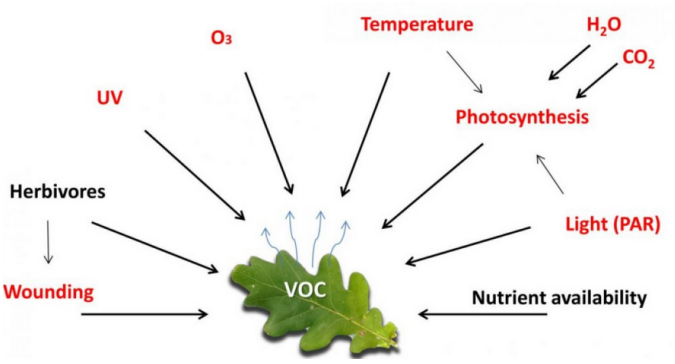


Fig. 1. VOCs and factors affecting plant emissions [3].

A typical E-nose system, as illustrated in Figure 2, consists of several key components including sensor arrays, signal transducers, and data processing units.

Also known as a spider plot, star plot, or Kiviat figure [5], the radar chart is more than just a graphical technique; it serves as a crucial methodological component in this study. It provides a straightforward way to display multivariate data on a two-dimensional plane, facilitating the visualization and comparison of sensor responses. This visualization ultimately enables the evaluation of sensor sensitivity [2].

A radar chart presents multivariate data on axes that radiate outward from a central point. As illustrated in Figure 3, each axis represents a distinct variable, with data points plotted along these axes. Connecting these points forms a polygon, and calculating the area of this polygon can yield valuable quantitative insights into the dataset [6].

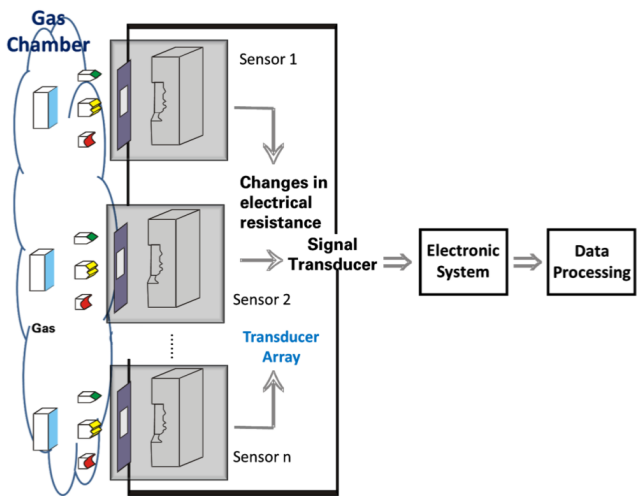


Fig. 2. Block diagram of an E-nose and its components, including sensors, signal transducer, electronic system, and data processing.

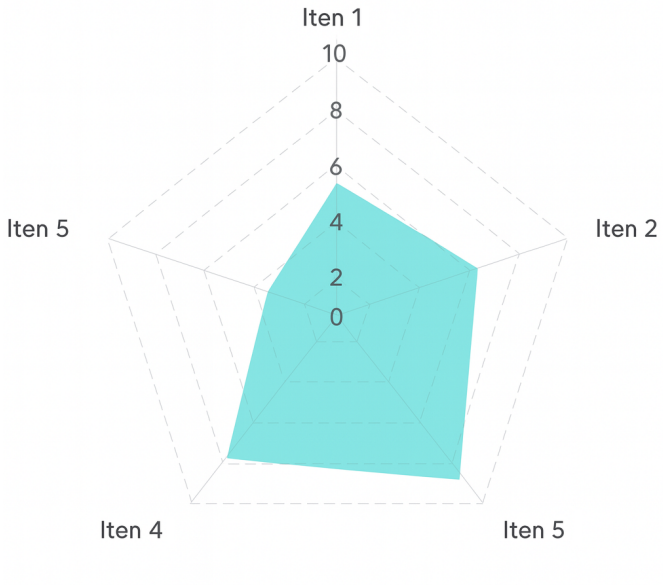


Fig. 3. Radar Chart with an Area.

A. Economic and Agricultural Context

Soybean cultivation occupies a strategic position in Brazilian agriculture [7]. This economic importance amplifies the impact of water stress, which can cause yield losses of up to 40% in soybean crops [8]. The development of affordable, non-invasive monitoring tools is therefore critical for sustainable agricultural management and food security.

Recent interdisciplinary collaborations combining engineering, physics, geosciences, plant sciences, ecophysiology, computer science, and instrumentation have advanced non-invasive phenotyping techniques [9]. Among these, E-nose technology offers particular promise due to its ability to detect VOCs emitted by plants under stress conditions in a non-invasive setup [10] [11].

B. VOC emissions throughout the day

Plant responses to water stress involve complex physiological and biochemical adaptations, including stomatal closure, growth repression, and activation of respiration pathways [12]. Recent studies have identified specific biomarkers, such as isoprene and hexanal, which serve as early, non-visible indicators that soybean plants are metabolically fighting to retain water before visible wilting occurs [13]. The relevance of compounds like hexanal is underscored by post-harvest studies, where it has been identified as a key marker of lipid peroxidation and oxidative deterioration in soybean grains, directly linked to stress-induced damage [14]. These VOC changes can precede visual symptoms by several days, providing a critical window for intervention.

Transpiration dynamics play a crucial role in these emissions, with Vapor Pressure Deficit (VPD) influencing stomatal behavior throughout the day [15]. Studies have shown distinct patterns between low VPD periods (7–11 AM) and high VPD periods (11 AM–3 PM), justifying our measurement protocol at 9:30 AM and 3:30 PM to capture these diurnal variations.

The ability to detect these subtle metabolic shifts through non-invasive methods like E-nose technology represents a significant advancement over traditional visual assessments. Our radar chart area methodology aims to capture these early VOC pattern changes, enabling proactive rather than reactive agricultural management.

”

1) Objectives and structure: Takenaka et al. [16] provided a method for evaluating the accessibility of a facility location using the area of a radar chart. The authors argue that the area of a radar chart is a more stable measure of accessibility than other metrics.

The main objectives of this investigation are:

1) To develop a method utilizing radar chart areas to evaluate the sensitivity of E-nose sensors in detecting water stress in soybean plants;
2) To identify which sensors within the E-nose array are most responsive to VOC changes associated with water stress;
3) To assess the effectiveness of radar chart areas as a quantitative metric for sensor performance in agricultural monitoring applications.

The remainder of this paper is structured as follows. Section III presents the materials and methods. Section IV discusses the experimental results. Finally, Section V concludes the article and outlines directions for future work.

II. State of the Art

This section presents a survey of relevant publications and investigations that situate our work within the current research landscape.

A. Bibliographic References for Radar Chart Area Calculation as Quantitative Result: a Survey

This section provides bibliographic references demonstrating that the calculation of graphical area in multidimensional radar charts serves as a quantitative result for performance measurement, evaluation, and comparison. The references are organized by methodology and application domain, with emphasis on papers that explicitly use area calculations as quantitative metrics.

1) Radar Chart Area Calculations Use:

a) Wind Power Accommodation Evaluation [17]: The authors proposed an improved radar chart method that replaced traditional fan-shaped sectors with quadrilateral evaluation regions. They constructed new area and perimeter vectors for the radar chart to provide comprehensive evaluation metrics. The area vector represented the aggregation degree of wind power accommodation ability, while the perimeter vector reflected the balance degree across different indicators.

This methodology allows for the evaluation of wind power systems based on both macroscopic indicators (e.g., installed capacity, power generation) and microscopic indicators (e.g., voltage stability, power quality).

b) Estimation Performance Evaluation [18]: The authors designed a comprehensive measure based on the radar chart's fan area and fan arc length, which they formalized into a radar chart index (RCI) that combines multiple performance measures through weighted components. Within this framework, the fan area served as the key quantitative component for calculating the overall estimation performance index. The proposed mathematical framework defines an index that integrates multiple otherwise incomprehensive measures, using fan area and arc length as core quantitative metrics, and is supported by case studies that demonstrate its utility through numerical comparisons. In application, this methodology enables the evaluation of estimation algorithms and statistical estimators across a range of performance criteria.

c) Principal Component Analysis Integration [19]: The authors combined PCA with radar charts to create a comprehensive evaluation model. Their key innovation was using the area of the radar chart polygon as a synthetic quantitative indicator, which they implemented by first applying PCA to transform and weight the original variables. The methodology involved applying PCA to reduce dimensionality, constructing a radar chart from the principal components, and then calculating the polygon area to serve as a comprehensive evaluation score, enabling ranking and comparison based on its magnitude. In application, this model facilitates general comprehensive evaluation and performance assessment across multiple domains.

d) Mathematical Foundations of Synthetic Indicators [20]: The authors established rigorous mathematical foundations for radar-chart-based synthetic indicators. They developed formal notation for radar-map-induced polygons, employed an analogue of the scalar product of vectors, proved theorems on polygon fields induced by radar maps, and constructed concentration indicators from radar-chart polygonal areas. Their key mathematical contributions include formally proving that radar chart polygons can serve as synthetic meters, constructing concentration measures (analogous to the Gini coefficient) from radar areas, and creating a theoretical framework for comparing alternatives using polygon-derived metrics. In application, this methodology is designed for socioeconomic indicator analysis, material deprivation studies, and country-level comparisons.

2) Radar Charts for Quantitative Shape Analysis:

a) Biomechanics and Materials Science [21]: The authors used permuted radar charts to create closed polygonal profiles representing multi-property mechanical performance and applied shape descriptors to these polygons for quantitative comparison. Specifically, they utilized polygon area as a measure of overall performance and shape metrics to compare performance distributions. This method is designed for application in comparative biomechanics, biological materials performance assessment, and functional morphology studies.

b) Construction Industry Performance [22]: The authors applied the radar chart to evaluate the performance of construction companies. Their quantitative method used radar chart areas to measure and compare operational performance across multiple dimensions, and they demonstrated how substituting traditional line or bar charts with area-based radar representations could provide holistic performance metrics.

3) Multi-Criteria Decision Making Applications:

a) Visual Filtering and Decision Support [23]: The authors constructed radial graphic representations with normalization to enable quantitative visual filtering of alternatives. Their system utilized radar-like plots with area-based comparisons to filter alternatives based on threshold criteria, provide quantitative filtering reductions, and support interactive multi-criteria selection. This approach was developed for application in multi-criteria decision support systems and alternative selection.

b) Educational Assessment [24]: The authors employed radar-like visual profiles for multi-competency comparison and quantitative assessment. Their tool used radar chart representations to provide numerical comparisons across student attributes and capabilities, making it applicable for student capability assessment and educational performance measurement.

4) Additional Supporting References:

a) Eco-efficiency Index Development [25]: The authors used radar charts as part of developing composite eco-efficiency indices, treating the visual representation as a quantitative tool for environmental performance comparison. Their methodology was designed for application in environmental performance assessment and sustainability metrics.

5) Mathematical Foundations and Area Calculation Methods:

a) General Polygon Area Formula: For a radar chart with n dimensions, where each dimension has a normalized value v_i (typically scaled 0-1 or 0-100) at angle $\theta_i = 2\pi i/n$, the

polygon area can be calculated as:

$$\text{Area} = \frac{1}{2} \sum_{i=1}^{n} (v_i \times v_{i+1} \times \sin(\theta_{i+1} - \theta_i)) \quad (1)$$

For equally-spaced axes (regular polygon case):

$$\text{Area} = \frac{n}{2} \times \sin\left(\frac{2\pi}{n}\right) \times \sum_{i=1}^{n} (v_i \times v_{i+1}) \quad (2)$$

where $v_{n+1} = v_1$ (closing the polygon).

b) Alternative Formulations: Several papers reference alternative area-based metrics:

- *Fan Area Method* [18]: Calculates area of individual sectors/fans and aggregates them with weights
- *Vector-Based Area* [17]: Constructs area vectors representing aggregation across multiple sub-regions
- *Normalized Area Index*: Area ratio comparing actual performance polygon to maximum possible area (all dimensions at maximum value)

6) Key Findings and Conclusions: The survey demonstrates that radar chart area calculations constitute a robust quantitative metric. This rigor stems from a formal mathematical foundation, including proofs and theorems that establish its theoretical validity. Its practical utility is corroborated by successful applications across diverse fields of study. Furthermore, the area metric provides distinctive comparative power, enabling objective numerical analysis among different alternatives. Ultimately, its principal virtue lies in its aggregation capability, efficiently synthesizing complex, multi-dimensional data into a single, interpretable quantitative indicator.

a) Application Domains: Radar chart area calculations have been validated as robust quantitative metrics through their successful application across diverse research domains. Within Engineering, they are utilized for evaluating wind power systems and estimation algorithms. In Biomechanics, these calculations assess the performance of biological materials. The field of Economics employs them to analyze socioeconomic indicators and derive concentration measures, while Education applies them for student capability assessment. Furthermore, in Business contexts, they facilitate the evaluation of construction company performance and eco-efficiency metrics. Finally, in Decision Science, radar chart areas provide foundational support for multi-criteria decision-making and systematic alternative selection.

b) Methodological Advantages: This survey highlights several distinct advantages of employing radar chart area as a quantitative metric, primarily its capacity for visual-quantitative integration, which merges intuitive graphical representation with objective numerical measurement. It further enables multi-dimensional synthesis by aggregating multiple disparate criteria into a single, comparable index. Crucially, the derived area exhibits sensitivity to balance, reflecting both the magnitude and the relative distribution of values across dimensions. The method demonstrates strong normalization compatibility, functioning effectively with standardized data, and provides comparative clarity by facilitating the straightforward ranking and benchmarking of alternatives. Collectively, this comprehensive bibliographic foundation firmly establishes radar chart area calculation as a quantitatively robust methodological approach, thereby justifying its application in this study to evaluate E-nose sensor sensitivity to water stress in soybean plants.

7) Recommended Core References: For researchers seeking to use radar chart area calculations as quantitative results, the core references summarized in Table I provide the strongest evidence across different domains.

III. Materials and Methods

Below are presented all the materials and methods used in this work, as well as the equations used to obtain the results.

A. Plant Material and Experimental Design

Soybean seeds (*Glycine max* L.) were germinated and grown in controlled greenhouse conditions at 25.0 ± 2.0 (°C), relative humidity of 60–70 (%), and a 14-hour photoperiod. Plants were cultivated in pots containing a standardized soil mixture and watered regularly to maintain optimal moisture levels.

Measurements were taken up to the V3 phenological stage of plant development. During the experiment (21 days), the plants were divided into two groups:

- **Irrigated (10 days):** Continued to receive regular irrigation to maintain field capacity.
- **Non-irrigated (11 days):** Subjected to water stress by withholding irrigation to reduce soil moisture content gradually.

The techniques and methods involve the direct manipulation of water availability, observation of dehydration symptoms (loss of turgor), quantification of growth response (biomass), and analysis of underlying physiological mechanisms (proteins and soluble sugars). Furthermore, environmental factors affecting water demand (such as humidity, a component of VPD) are considered crucial for accurate modeling of water stress. [26] [27]

B. Electronic Nose Setup

The E-nose system used was an Alpha Fox™ 2000, equipped with an array of six Complementary Metal-Oxide Semiconductor (CMOS) sensors. These sensors operate on the metal-oxide semiconductor principle illustrated in Figure 4, where gas adsorption/desorption processes induce measurable resistance changes through electron transfer mechanisms.

The specific equipment used in this study, shown in Figure 5, features a controlled sample injection system and integrated data acquisition interface, enabling precise VOC measurements under standardized conditions.

The Alpha Fox system employed a six-sensor array with specific sensitivities to different VOC classes, as detailed in Table II.

TABLE I
CORE REFERENCES FOR RADAR CHART AREA AS QUANTITATIVE METRIC

Category	Author	Key Contribution
Mathematical Foundation	Borkowski et al. [20]	Formal proofs for radar chart polygons as synthetic meters
Engineering	Peng et al. [18]	Radar Chart Index (RCI) with area metrics
	Li et al. [17]	Area/perimeter vectors for wind power evaluation
Biomechanics	Porter & Niksiar [21]	Polygon area for biological materials performance
Statistical Methods	Wang et al. [19]	PCA integration with polygon area indicator

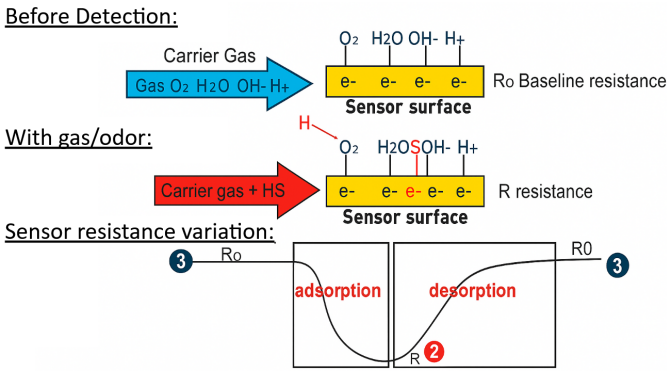


Fig. 4. Operating principle of metal-oxide semiconductor (MOS) sensors showing adsorption/desorption processes and corresponding resistance changes. Red arrows indicate electron transfer during gas detection [29].

Fig. 5. Commercial Alpha Fox^{TM} 2000 electronic nose system used in experiments [29].

TABLE II
THE SENSORS INSTALLED IN THE E-NOSE AREA [28].

No.	Sensor	Sensitivity property	Reference Materials
S1	T30/1	Organic compounds	Organic compounds
S2	P10/1	Combustible gas	Hydrocarbon
S3	P10/2	Inflammable gas	Methane
S4	P40/1	Oxidizing gas	Fluorine
S5	T70/2	Aromatic compounds	Methyl benzene, xylene
S6	PA/2	Organic compounds and toxic gas	Ammonia, amines, ethyl alcohol

C. Chamber Design and Instrumentation

The experimental setup included a specially designed chamber (Figure 6) with dimensions d=25.0 cm, height=57.0 cm, and volume=27.93 L, equipped with complementary sensors for monitoring environmental parameters [2]. The chamber, constructed from Poly(methyl methacrylate) (PMMA) with 92% transmittance in the visible range [30], was instrumented to measure:

- Temperature (T in °C) using digital thermometers with 0.1°C resolution
- Relative humidity (RH in %) with 0.5% resolution sensors
- CO_2 concentration (0–2,000 ppm range) using Vaisala probes
- Natural light intensity (0.001–19.9 K lumen/m^2)

Continuous monitoring was performed at 5-minute intervals, creating a comprehensive environmental database alongside E-nose measurements [2]. The chamber design included a computer fan to simulate wind (28.32 L/min) and an irrigation system that allowed water administration without compromising chamber insulation.

D. Soil Specifications and Isolation

The experiment utilized dystrophic Red Yellow Latosol (LVAd) with specific granulometry: 369 g/kg clay, 54 g/kg silt, and 577 g/kg sand. Soil moisture was maintained at field capacity (0.295 cm^3/cm^3 at 10 kPa) during irrigation phases and reached the permanent wilting point (0.134 cm^3/cm^3 at -1,500 kPa) during stress phases.

To isolate transpiration effects, the soil was covered with aluminum foil, effectively eliminating gas exchange between the rhizosphere and the chamber atmosphere. This isolation method, consistent with approaches validated for soybean transpiration studies [31], ensured that measured VOC emissions originated primarily from plant physiological processes rather than soil microbial activity.

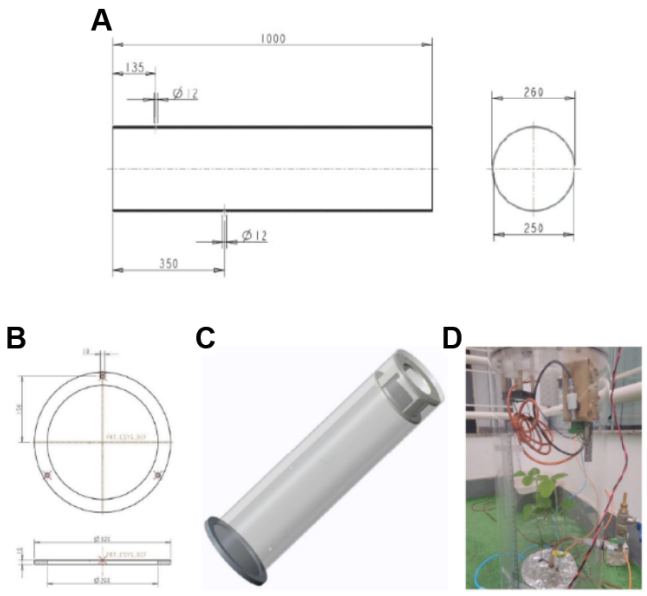


Fig. 6. Experimental chamber design for soybean VOC monitoring. Key specifications: 250 mm internal diameter × 570 mm height (27.93 L volume). Components shown: (A) Dimensional schematic, (B) Base plate with sensor/irrigation interfaces, (C) Assembly diagram, (D) Operational setup with plant specimen.

E. Gas Sampling Using the Headspace Protocol

The headspace sampling technique using a 2,500 µL syringe with PTFE seal provided precision of ¡ ±1% volume accuracy. Baseline measurements from the empty chamber over three days established reference conditions: temperature variation of 4.0°C (23.0–27.0°C), relative humidity variation of 9.0% (16–25%), and CO_2 variation of 20.0 ppmv (250–270 ppmv). These controlled baseline measurements ensured that subsequent plant-emitted VOC detections were not confounded by chamber artifacts.

The sampling volume of 500 µL at a flow rate of 150 mL/min was optimized to balance signal intensity with chamber disturbance minimization, based on preliminary sensitivity tests.

F. Applications in Electronic Nose

The sensitivity S (%) for each sensor was calculated using Equation (3):

$$S(\%) = \left(\frac{R - R_0}{R_0}\right) \times 100\,(\%) \tag{3}$$

where:

- R_0 – Initial electrical resistance (Ω);
- R – Electrical resistance varying over time (Ω).

To analyze the data obtained from the E-nose, we utilized both radar charts and radar area charts to represent the peak sensitivity (S (%)) for each of the six sensors: S1 T30/1, S2 P10/1, S3 P10/2, S4 P40/1, S5 T70/2, and S6 PA/2. This data was normalized using Equation (3) and is shown in Figure 7.

A radar area chart is a specific type of radar chart that illustrates the values by displaying the area enclosed by the lines connecting the data points. Figure 8 presents the representation of both the radar chart and the radar area chart for the peak sensitivity (S (%)) across all sensors. Radar charts are particularly useful for visualizing multiple variables simultaneously [32].

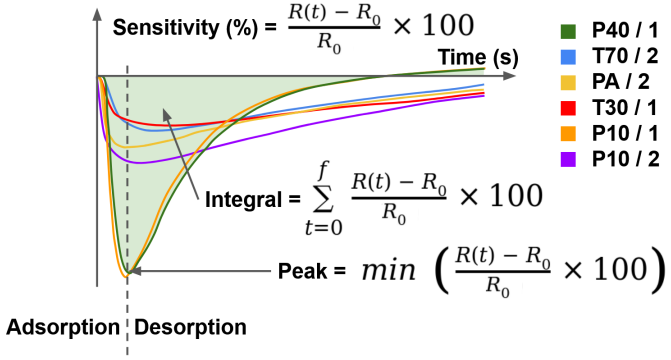


Fig. 7. The variation in sensitivity, using Equation (3), of each of the six sensors in relation to time, depending on the gas sampled and measured in the E-nose.

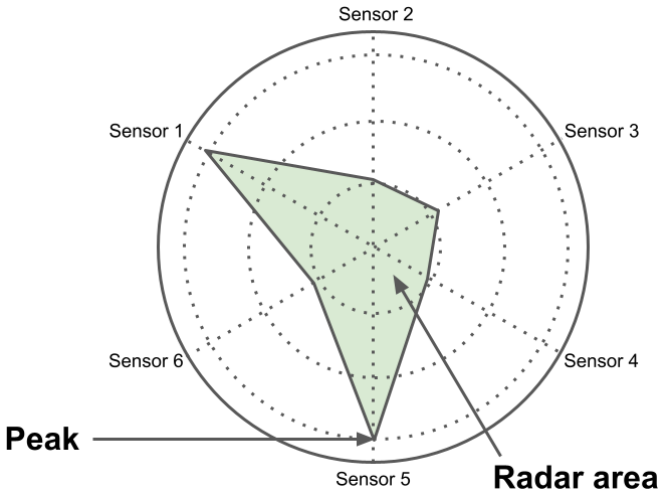


Fig. 8. The radar chart and the radar area chart from the sensitivity (%) peak for the six sensors (S1: T30/1; S2: P10/1; S3: P10/2; S4: P40/1; S5: T70/2; and S6: PA/2) from the E-nose.

G. Data Processing from E-Nose Responses

1) Data Processing from E-Nose Responses: The transformation of raw sensor signals into quantitative metrics followed the workflow depicted in Figure 9. This systematic approach ensured consistent feature extraction across all measurements.

The raw E-nose data, consisting of resistance-time curves for each of the six sensors during 240-second adsorption/desorption cycles, was processed to extract quantitative features. For each sensor response curve $r_i(t)$, where $i = 1, \ldots, 6$ represents the sensor index and t represents time in seconds, two key metrics were computed (Figures 7 and 8):

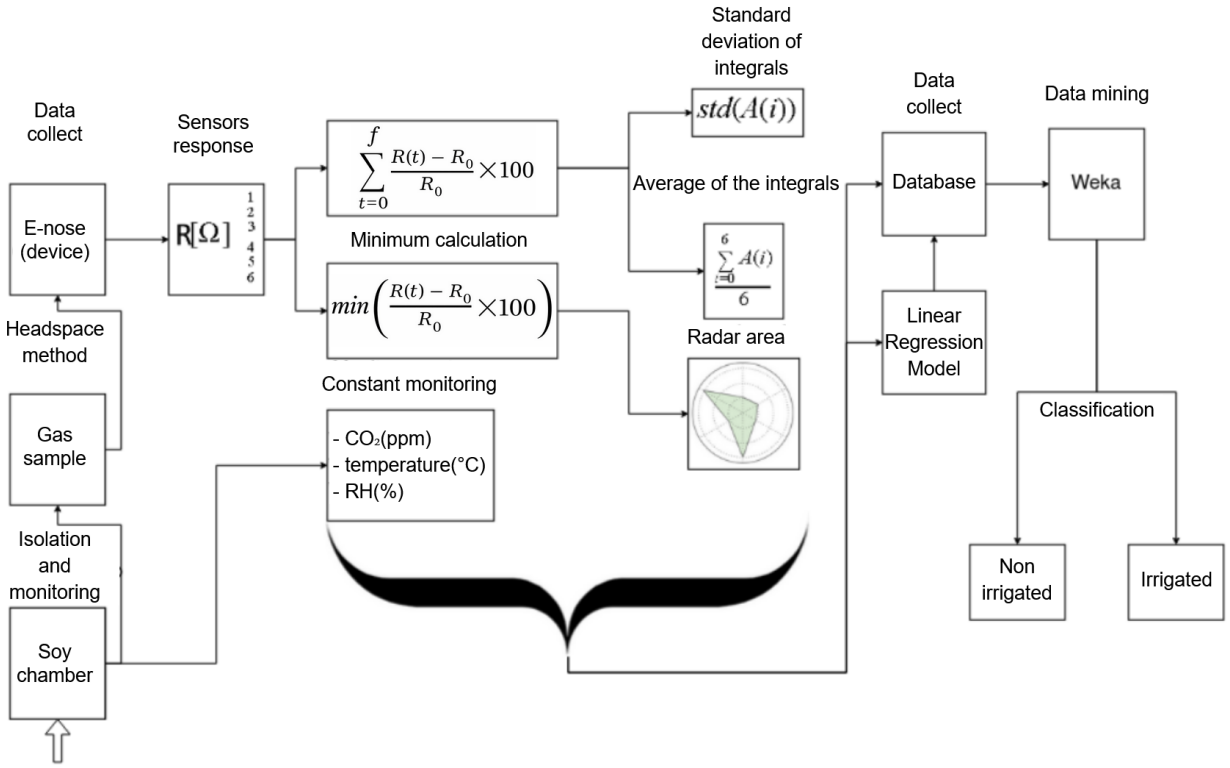


Fig. 9. Data processing workflow from raw sensor responses to derived metrics. Steps include: (1) Raw resistance-time curves for six sensors, (2) Calculated sensitivity profiles, (3) Peak detection and area calculation, (4) Radar chart construction from normalized peaks.

- **Area under the curve:** $A_i = \sum_{t=0}^{240} |\frac{r_i(t)-r_i(0)}{r_i(0)}|$, representing the cumulative response intensity over the measurement period
- **Peak sensitivity:** $P_i = \min_{t\in[0,240]} \frac{r_i(t)-r_i(0)}{r_i(0)} \times 100\%$, representing the maximum response amplitude

These metrics were computed for all three replicates (3 × 500 µL samples) taken during each measurement session (morning and afternoon). The radar chart area A_n was subsequently calculated from the normalized peak sensitivity values P_i of all six sensors, providing a composite measure of overall sensor array response.

The processing pipeline ensured robust feature extraction by:

1) Computing baseline resistance $r_i(0)$ from stable pre-injection measurements
2) Normalizing responses across sensors to account for differential sensitivities
3) Averaging triplicate measurements to reduce sampling variability
4) Validating data quality through response curve shape analysis

H. Experimental Protocol Refinements

The methodology evolved through three experimental phases, as summarized in Table III. These iterative refinements enhanced measurement reliability while addressing technical challenges identified in earlier trials.

The study incorporated iterative protocol improvements across multiple experimental iterations with six soybean specimens:

- **Sample volume optimization:** Increased from single 500 µL samples to 1,500 µL samples (3 × 500 µL replicates) for improved statistical robustness and outlier detection
- **Temporal coverage:** Expanded from single daily measurements to both morning (9:30 AM) and afternoon (3:30 PM) sessions to capture diurnal variability
- **Environmental controls:** Implemented dry air purging systems in later experiments to maintain consistent chamber humidity levels
- **Extended monitoring:** Prolonged irrigation periods for selected specimens to decouple plant developmental age from water stress effects

These refinements, developed through systematic experimentation, enhanced data quality and enabled more robust pattern recognition in subsequent analyses while maintaining the non-invasive nature of the measurements.

I. Calculating the Area of a Radar Chart

A radar chart is a graphical representation that effectively illustrates multidimensional data by expressing the values of each attribute in a clear and concise manner. Its 2D visualization provides a comprehensive view of the data, making it easier to analyze and understand its various dimensions [33].

The method of radar chart for Multidimensional Data: $X = \{X_1, X_2, \ldots, X_j, \cdots X_n\}$ is a multi-dimensional data set, and $X_i = \{x_{i1}, x_{i2}, x_{i3}, \ldots, x_{iN}\}$ is a N-dimensional vector. Use the radar chart when $N \geqslant 3$ [33].

A method for evaluating the accessibility of a facility location using the area of a radar chart was provided by Takenaka et al. [16]. The authors argue that the area of a radar chart is a more stable measure of accessibility than other measures.

The radar values were calculated using the absolute value of the minimum values in Equation (3) for each sensor, where:

$$X_i = S_i = \{S_1(\%), S_2(\%), S_3(\%), S_4(\%), S_5(\%), S_6(\%)\} \tag{4}$$

To calculate the area (A_n) of the polygon formed in a radar chart:

1) **Convert Polar Coordinates to Cartesian Coordinates.** Each data point is defined by:
   - r_i: The distance from the center to the data point along axis i (the normalized value of the variable).
   - θ_i: The angle corresponding to axis i .

   The Cartesian coordinates (x_i, y_i) are related to the polar coordinates (r_i, θ_i) by the formulas:

$$x_i = r_i \cos\theta_i, \quad y_i = r_i \sin\theta_i \tag{5}$$

2) **Apply the Shoelace Formula:** The area of the polygon can be calculated using the Shoelace Formula (6):

$$A_n = \frac{1}{2}\left|\sum_{i=1}^{n}(x_i y_{i+1} - x_{i+1} y_i)\right| \quad (\%^2) \tag{6}$$

where $x_{n+1} = x_1$ and $y_{n+1} = y_1$ to complete the loop.

The formula sums the cross-products of vertex coordinates in a specific order [2].

J. S-Norm: Geometric Interpretation of the Radar Area

The area of the polygon is crucial because the square of the vector's S-norm is equal (minus a multiplicative constant) to the area of the polygon formed by the radar graph of these vectors.

a) Definition of the S-Scalar Product and the S-Norm: The formalism relies on the concept of the S-operator (S), which is an *S-shift* (displacement) operator defined for a vector $x = (x_1, x_2, \ldots, x_n) \in X = \mathbb{R}^n$ as $Sx := (x_2, x_3, \ldots, x_n, x_1)$.

The S-scalar product is defined by the formula:

$$\langle x, y\rangle_S := \langle x, Sy\rangle \tag{7}$$

where $\langle\cdot\rangle$ represents the standard scalar product.

The S-norm of a vector x (denoted $\|x\|_S$) is defined by Formula (8) as:

$$\|x\|_S := |\langle x, x\rangle_S|^{1/2} = \left(\frac{1}{n}\left|\sum_{k=1}^{n} x_k x_{k+1}\right|\right)^{1/2} \tag{8}$$

where $x = (x_1, x_2, \ldots, x_n)$, and x_{n+1} is defined as x_1.

b) Geometric Interpretation and Area Formula: Assuming that the vector $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ has coordinates $1 \geq x_i \geq 0$, the graphical representation utilizes an n-regular polygon inscribed in a unit circle (radius $r = 1$) with the center at the origin. The coordinates x_i are represented as points x_i on the 0_i axes emanating from the center.

By connecting the points in order (x_1 to x_2, x_2 to x_3, …, x_n to x_1), an n-polygon is obtained.

The area (S_1) of this polygon (induced by vector x) is defined by the formula:

$$S_1 = \sum_{i=1}^{n} \frac{1}{2} x_i x_{i+1} \sin\left(\frac{2\pi}{n}\right) \tag{9}$$

This formula can be rewritten in terms of the squared S-norm:

$$S_1 = \frac{1}{2}\sin\left(\frac{2\pi}{n}\right)\sum_{i=1}^{n} x_i x_{i+1} = \frac{1}{2} n \sin\left(\frac{2\pi}{n}\right)\|x\|_S^2 \tag{10}$$

where $x_{n+1} := x_1$.

IV. Results and Discussion

The average values and standard deviations derived from radar measurements of both irrigated and non-irrigated soybean samples are illustrated in Figure 10.

A. Temporal Variability and Diurnal Patterns

TABLE III
Experimental iterations and protocol refinements across six soybean specimens

Specimen	Sampling Protocol	Environmental Controls	Key Refinements
Plants 1-3	Single 500 µL samples	Basic chamber monitoring	Baseline methodology establishment
Plant 4	3 × 500 µL replicates	Dry air purging implemented	Improved statistical robustness
Plants 5-6	Morning/afternoon sessions	Open chamber periods	Diurnal pattern capture

The experimental design captured distinct diurnal patterns in VOC emissions, with measurements taken at 9:30 AM (morning) and 3:30 PM (afternoon) [2]. Morning measurements consistently showed lower standard deviations in radar area calculations and more stable sensor response patterns. In contrast, afternoon measurements exhibited increased variability, particularly in response to environmental fluctuations such as light intensity changes.

This temporal variability is evident in Figure 10, where afternoon sessions show greater standard deviations. The most pronounced variation occurred on the 22nd day, characterized by overcast conditions with intermittent rain, resulting in luminosity variations of approximately 77% during afternoon measurements compared to 39% in the morning [2].

TABLE IV
MACHINE LEARNING CLASSIFICATION PERFORMANCE FOR WATER STRESS DETECTION.

Algorithm	Major Accuracy Obtained (%)
Decision Tree	93.6
k-NN (k=3)	80.7
Logistic Regression	78.2

B. Environmental Parameter Correlations

Complementary environmental measurements revealed significant patterns [2]:

- Chamber temperature consistently exceeded laboratory environment by 2–4°C, with maximal differences observed in afternoon measurements.
- CO_2 concentrations inside the chamber were typically 50–100 ppm lower than external levels during active photosynthesis periods.
- Relative humidity within the chamber exceeded external levels by 15–25% during morning measurements.

These environmental interactions, consistent with plant physiological responses to water stress [34], [35], contextualize the E-nose responses and highlight the importance of standardized measurement conditions for agricultural applications.

C. Synthesis with Machine Learning Validation

The radar area methodology developed in this study aligns with and complements machine learning approaches applied to the same experimental system. Comparative performance analysis of three machine learning algorithms (Table IV) achieved 93.6% classification accuracy using Decision Tree algorithms on comprehensive datasets incorporating both E-nose responses and environmental parameters [36].

1) Machine Learning Pipeline Implementation: A comprehensive machine learning pipeline was implemented using the WEKA toolkit (Figure 11), incorporating:

1) **Data balancing:** Equalizing irrigated and non-irrigated samples to prevent classifier bias
2) **Feature normalization:** Scaling all numerical features to [0,1] range for algorithm compatibility
3) **Algorithm evaluation:** Comparative testing of Decision Trees, k-Nearest Neighbors (KNN), and Logistic Regression
4) **Validation strategies:** Three-fold validation using training set evaluation, cross-validation, and independent test set validation

D. Analysis of Measurement Variability

The data reveal a marked difference between the measurements taken in the morning and those taken in the afternoon, with the greatest standard deviation observed during the afternoon sessions.

This pronounced variation may be attributed to several influencing factors, including the physiological state of the plants, which can change due to water uptake and nutrient availability.

Environmental conditions at the time of sample extraction also likely played a significant role; the fluctuating temperatures, humidity levels, and light intensity throughout the day can affect the plants' responses. Furthermore, the specific growth stage of the soybeans—whether they are in vegetative growth or nearing maturation—can impact how they interact with their environment. Additionally, potential errors in the syringe headspace during sampling could introduce variability in the measurements.

It is particularly noteworthy that the highest standard deviation was recorded during the afternoon. On the 22nd day of the experiment, specific weather conditions were present, characterized by overcast skies, intermittent rain, and significant cloud cover. These factors likely influenced the plants' physiological responses, contributing to the observed variability in the data.

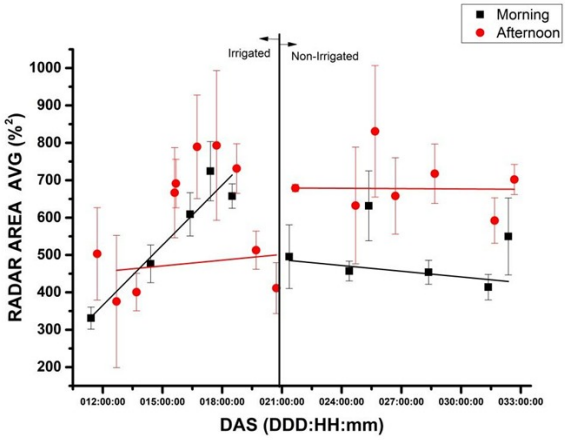


Fig. 10. E-nose measurements of gas samples taken from a chamber containing soybeans during the Days After Sowing (DAS), using the average radar area and standard deviation ($n = 3$). The measurements are presented based on the time of day, either in the morning (9:30 a.m.) denoted by red circles or in the afternoon (3:30 p.m.) denoted by black squares. Moreover, the measurements are obtained from both irrigated and non-irrigated plants. For each DAS, gas samples are measured three times in both periods, i.e., the morning and afternoon, to obtain the radar area measurement.

E. Sensor Performance Analysis

Individual sensor analysis revealed differential sensitivity patterns: sensors P10/1 (combustible gas/hydrocarbon) and P40/1 (oxidizing gas/fluorine) showed the highest responsiveness to water stress-induced VOC changes. During irrigated conditions (DAS 11–21), peak sensitivity for P10/1 was -27.97% ± 4.36%, while during water stress (DAS 22–32) it was -28.62% ± 3.26%. For P40/1, corresponding values were -28.30% ± 4.87% and -28.88% ± 3.59%.

The negative sensitivity values indicate decreased electrical resistance upon gas exposure, reflecting increased sensor conductivity in response to specific VOC compounds emitted by water-stressed plants.

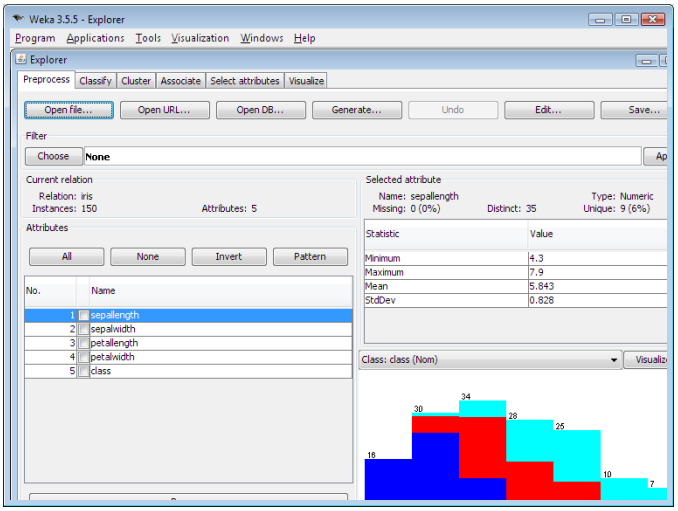


Fig. 11. WEKA (Waikato Environment for Knowledge Analysis) software interface showing the machine learning workflow [37].

F. Methodological Validation and Limitations

The experimental approach addressed several methodological challenges:

- Chamber design minimized environmental interference while maintaining plant viability
- Soil isolation controlled for non-plant VOC sources
- Temporal standardization reduced diurnal variability effects
- Multiple measurement replicates (n=3 per time point) ensured statistical robustness

Notably, the 22nd day of experimentation presented unique conditions with overcast skies and intermittent rain, resulting in the highest observed standard deviations. These conditions provided valuable insight into environmental sensitivity and highlight the importance of meteorological considerations in field applications.

V. Conclusion and Future Work

The radar area chart is a specialized variant of the radar chart that utilizes the area enclosed by the connecting lines of data points to visually represent and compare values. Figure 8 illustrates this radar chart format, specifically highlighting the radar area at the sensitivity peak (S (%)), which indicates the maximum responsiveness of the variables in question.

Radar area charts are particularly valuable tools when analyzing and comparing the overall performance of distinct data groups, for example, across different experimental conditions. This visual representation facilitates the interpretation of complex multivariate sensor data, enabling researchers to quickly identify patterns and trends relevant to plant stress detection. The area can be used as a valid metric to rank data.

The key findings demonstrate that:

- The radar chart area metric effectively distinguished irrigated from non-irrigated soybean plants, showing a clear divergence after the onset of water stress.
- Sensor-specific analysis identified S2 (P10/1) and S4 (P40/1) as the most responsive to water-stress-induced VOC changes.
- Measurements exhibited significant diurnal variability, with morning sessions (9:30 AM) providing more stable and consistent radar area values compared to afternoon sessions (3:30 PM), underscoring the importance of standardized sampling times.
- The geometric interpretation of the area, supported by the mathematical formalism of the S-norm, provides a solid theoretical foundation for using polygon area as a synthetic indicator of overall sensor array response.

The integration of this area-based metric with machine learning validation (achieving up to 93.6% classification accuracy) confirms its utility as a discriminative feature for automated stress detection systems. By capturing early, non-visible VOC pattern shifts, this methodology enables a proactive approach to irrigation management, potentially preventing yield losses before visual symptoms appear.

Future work directions include integrating the method with equipment in a mobile unit to facilitate field use and applying the methodology to study thermal and water stress.

Initial studies with wheat, aimed at investigating water stress, are being carried out using the proposed technique and methodology, and are not limited to soybeans.

Acknowledgment

This work was supported by Embrapa Instrumentation and Embrapa project # 20.22.01.001.00. The authors extend their gratitude to Dr. Ednaldo José Ferreira from Embrapa Instrumentation, the Embrapa-Bayer project 10.20.05.006.00.00, and the teams at LANAPRE and LNNA for providing experimental facilities and materials. Additional acknowledgments are due to CAPES-CNPq for doctoral fellowships and to ICMC-USP for institutional support.

References

[1] P. S. d. P. Herrmann and M. S. Lucas, "Using radar chart areas to evaluate the sensitivity of electronic nose sensors in detecting water stress in soybean," *ALLSENSORS*, 2025, thinkMind Digital Library, https://www.thinkmind.org.

[2] P. S. D. P. Herrmann, M. d. Santos Luccas, E. J. Ferreira, and A. Torre Neto, "Application of electronic nose and machine learning used to detect soybean gases under water stress and variability throughout the daytime," *Frontiers in Plant Science*, vol. 15, p. 1323296, 2024.

[3] D. Materić, D. Bruhn, C. Turner, G. Morgan, N. Mason, and V. Gauci, "Methods in plant foliar volatile organic compounds research," *Applications in Plant Sciences*, vol. 3, no. 12, p. 1500044, 2015.

[4] S. Zorpeykar, E. Mirzaee-Ghaleh, H. Karami, Z. Ramedani, and A. D. Wilson, "Electronic nose analysis and statistical methods for investigating volatile organic compounds and yield of mint essential oils obtained by hydrodistillation," *Chemosensors*, vol. 10, no. 11, p. 486, 2022.

[5] K. W. Kolence and P. J. Kiviat, "Software unit profiles & kiviat figures," *ACM SIGMETRICS Performance Evaluation Review*, vol. 2, no. 3, pp. 2–12, 1973.

[6] M. J. Saary, "Radar plots: a useful way for presenting multivariate health care data," *Journal of Clinical Epidemiology*, vol. 61, no. 4, pp. 311–317, 2008.

[7] Embrapa Soja, "Soja em números," https://www.embrapa.br/en/soja/cultivos/soja1/dados-economicos, 2025, accessed in: 15 Dec. 2025.

[8] O. Basal and A. Szabó, "Physiology, yield and quality of soybean as affected by drought stress," Tech. Rep., 2020.

[9] F. Fiorani and U. Schurr, "Future scenarios for plant phenotyping," *Annual Review of Plant Biology*, vol. 64, no. 1, pp. 267–291, 2013.
[10] B. Niederbacher, J. Winkler, and J. Schnitzler, "Volatile organic compounds as non-invasive markers for plant phenotyping," *Journal of Experimental Botany*, vol. 66, no. 18, pp. 5403–5416, 2015.
[11] R. Jansen, J. Wildt, I. Kappers, H. Bouwmeester, J. Hofstee, and E. Van Henten, "Detection of diseased plants by analysis of volatile organic compound emission," *Annual Review of Phytopathology*, vol. 49, no. 1, pp. 157–174, 2011.
[12] M. G. Hale and D. M. Orcutt, *The Physiology of Plants under Stress*. New York, USA: Wiley, 1987.
[13] M. G. Mostofa, A. Sahu, Y. Xu, I. Basrai, L. Doron, V. Lefrancois, and T. D. Sharkey, "Cryptic isoprene emission of soybeans," *Proceedings of the National Academy of Sciences*, vol. 122, no. 24, p. e2502360122, 2025.
[14] V. Ludwig, M. R. P. Berghetti, S. R. Ribeiro, F. P. Rossato, L. M. Wendt, F. R. Thewes, F. R. Thewes, A. Brackmann, V. Both, and R. Wagner, "The effects of soybean storage under controlled atmosphere at different temperatures on lipid oxidation and volatile compounds profile," *Food Research International*, vol. 147, p. 110483, 2021.
[15] T. R. Sinclair, C. D. Messina, A. Beatty, and M. Samples, "Assessment across the united states of the benefits of altered soybean drought traits," *Agronomy Journal*, vol. 102, no. 2, pp. 475–482, 2010.
[16] T. Takenaka, K. Nakamura, T. Ukai, and Y. Ohsawa, "Stability of the area of radar chart to evaluate the accessibility of facility location," *J. City Plann. Institute Japan*, vol. 53, no. 3, pp. 640–645, 2018.
[17] G. Li, G. Li, and M. Zhou, "Comprehensive evaluation model of wind power accommodation ability based on macroscopic and microscopic indicators," *Protection and Control of Modern Power Systems*, vol. 4, no. 3, pp. 1–12, 2019.
[18] W. Peng, Y. Li, Y. Fang, Y. Wu, and Q. Li, "Radar chart for estimation performance evaluation," *IEEE Access*, vol. 7, pp. 113 880–113 888, 2019.
[19] D. Wang, Y. Wan, X. Wang *et al.*, "Improved radar chart based on principal component and its application in comprehensive evaluating," *J. Appl. Stat. Manag*, vol. 29, no. 5, pp. 883–889, 2010.
[20] B. Borkowski, A. Wiliński, W. Szczesny, and Z. Binderman, "Mathematical analysis of synthetic measures based on radar charts," *Mathematical Modelling and Analysis*, vol. 25, no. 3, pp. 473–489, 2020.
[21] M. M. Porter and P. Niksiar, "Multidimensional mechanics: Performance mapping of natural biological systems using permutated radar charts," *PloS one*, vol. 13, no. 9, p. e0204309, 2018.
[22] A. Reske Filho, "Aplicação do gráfico radar na avaliação do desempenho das empresas de construção civil," Ph.D. dissertation, 2014.
[23] A. A. Zakharova, D. A. Korostelyov, and O. N. Fedonin, "Mathematical support and software of visual filtering of alternatives in multi-criteria decision making problems," in *CEUR Workshop Proceedings*. CEUR-WS.org, 2019, pp. 82–85.
[24] F. Ye, Y. Chen, and Q. Huang, "A study of the visualization tool for computer science majors' capability assessment," *International Journal of Information and Education Technology*, vol. 9, no. 1, pp. 61–65, 2019.
[25] C. P. Pereira, D. P. Paes, D. M. Prata, and L. P. Monteiro, "Desenvolvimento de índice de comparação de ecoeficiência a partir de ecoindicadores," *Sistemas & Gestão*, vol. 9, no. 2, pp. 168–180, 2014.
[26] R. S. Kong and H. A. Henry, "Interactions of plant growth responses to spring freezing and summer drought: a multispecies comparison," *American Journal of Botany*, vol. 106, no. 4, pp. 531–539, 2019.
[27] E. Kirchhof, F. Campos-Arguedas, N. S. Arias, and A. P. Kovaleski, "Thresholds for spring freeze: measuring risk to improve predictions in a warming world," *New Phytologist*, vol. 248, no. 2, pp. 563–575, 2025.
[28] W. Wei, J. Li, and L. Huang, "Discrimination of producing areas of astragalus membranaceus using electronic nose and uhplc-pda combined with chemometrics." *Czech Journal of Food Sciences*, vol. 35, no. 1, 2017.
[29] FOX Analyzer, *Hardware User's Guide – Manuel Number 001*, 2000, alpha Fox^{TM} 2000 Electronic Nose System.
[30] P. E. Keller and R. T. Kouzes, "Water vapor permeation in plastics," no. PNNL-26070, 2017. [Online]. Available: https://www.osti.gov/servlets/purl/1411940
[31] L. C. Ferreira, W. Neiverth, L. F. F. Maronezzi, R. N. R. Sibaldelli, A. L. Nepomuceno, J. R. B. Farias, and N. Neumaier, "Efficiency of cover materials in preventing evaporation in drought-stressed soybeans grown in pots." *Revista de Ciências Agrárias*, vol. 58, no. 4, pp. 359–365, 2015.
[32] W.-Y. Liu, B.-W. Wang, J.-X. Yu, F. Li, S.-X. Wang, and W.-X. Hong, "Visualization classification method of multi-dimensional data based on radar chart mapping," in *2008 International Conference on Machine Learning and Cybernetics*, vol. 2. IEEE, 2008, pp. 857–862.
[33] W. Peng, "Improved radar chart for lighting system scheme selection," *Applied Optics*, vol. 61, no. 19, pp. 5619–5625, 2022.
[34] H. Lambers and R. S. Oliveira, "Plant water relations," in *Plant Physiological Ecology*. Springer, 2019, pp. 187–263.
[35] J. Pallas Jr, "Transpiration and stomatal opening with changes in carbon dioxide content of the air," *Science*, vol. 147, no. 3654, pp. 171–173, 1965.
[36] P. S. P. Herrmann and M. Santos Luccas, "Utilização do "e-nose" e "machine learning" para investigação da emissão de gases da soja submetida a estresse hídrico," Embrapa Instrumentação, Relatório Técnico PIBIC/CNPq Processo CNPq 129877/2019-0, 2020.
[37] Wikipedia contributors, "Weka (software)," https://pt.wikipedia.org/wiki/Weka, 2025, accessed in: 15 Dec. 2025.

Coordinates Are Just Features: A Benchmark Study of Geospatial Modeling

1st Yameng Guo
dept. Business Informatics and Operations Management
Ghent University
Gent, Belgium
0000-0003-2719-1356
yameng.guo@ugent.be

2nd Seppe vanden Broucke
dept. Business Informatics and Operations Management
Ghent University
Gent, Belgium
Research Centre for Information Systems Engineering
KU Leuven
Leuven, Belgium
0000-0002-8781-3906
seppe.vandenbroucke@ugent.be

***Abstract*—Geospatial inference has long been recognized as a critical topic of research. Modeling approaches in this area can generally be categorized into two main types, i.e., explicit and implicit spatial dependence learning. The key difference between these categories lies in whether spatial information (typically coordinates) is used as input to a distance function or simply treated as standard features in a machine learning algorithm. Traditional geospatial statistical models, such as Geographically Weighted Regression and Kriging, explicitly model spatial dependence. However, they often suffer from high computational costs and struggle to balance the trade-off between predictive performance and efficiency. In this work, we aim to demonstrate that explicitly modeling geospatial dependence is often not necessary. Treating coordinates as standard input features can yield competitive predictive performance while significantly reducing computational overhead, provided that a sufficiently capable learner is used. To substantiate our claims, we conduct an extensive comparison across a wide range of models. As an extended version of our previous work, we broaden the scope of models considered and include additional tabular deep learning models based on the transformer architecture and its attention mechanism. We also assess the statistical significance of performance differences across datasets. Furthermore, we include an interpretability analysis to examine the role of coordinates in models that learn spatial information either explicitly or implicitly. Our results show that even models, which treat coordinates as standard features can achieve competitive performance, with substantially lower training costs, while still effectively capturing spatial dependence. To the best of our knowledge, this is the first comprehensive study to evaluate both the effectiveness and efficiency of using coordinate inputs directly in spatial prediction tasks across a diverse set of modeling paradigms.**

***Keywords-geospatial regression*; *tabular deep learning*; *ensembles modeling*; *spatial statistics*; *comparative performance*.**

I. Introduction

Spatial inference [1] plays a critical role across various industries, including environmental science [2] [3], urban planning [4], and disaster management [5] [6], where predicting unobserved values at specific locations is essential for informed decision-making in real-world applications.

To improve geospatial inference performance, researchers have developed numerous approaches that leverage both spatial and non-spatial information. Broadly, these approaches fall into two categories: explicit spatial dependence learning and implicit spatial dependence learning. Explicit spatial dependence learning relies on the principle that geographically closer observations are more likely to be similar. It typically treats location information (e.g., coordinates) as separate inputs and fits a distance function to estimate the influence of nearby points on a target location. Prominent methods such as Kriging [7] [8] and Geographically Weighted Regression (GWR) [9] embody this principle using variograms or distance-decay weighting, offering both interpretability and predictive power, and have been widely adopted for spatial interpolation and regression tasks.

In contrast, implicit spatial dependence learning does not explicitly estimate distance functions. Instead, it treats coordinates as regular features, integrating them into the model alongside other non-spatial features. This category includes traditional Machine Learning (ML) models such as linear regressors, Gaussian Processes (GPs), and tree ensembles, which excel at capturing nonlinear relationships while offering training efficiency.

Beyond classical ML, Tabular Deep Learning (TDL) has gained significant attention with the rise of deep learning. Transformers, in particular, provide powerful solutions for tabular tasks due to their flexibility and ability to model complex interactions among heterogeneous features. For example, the Prior-Data Fitted Network (PFN) Transformer [10] enables efficient supervised learning on small datasets without the need for additional hyperparameter tuning. Other deep learning models such as Neural Oblivious Decision Ensembles (NODE) [11] and Gated Additive Tree Ensemble (GATE) [12] have also shown promising performance in tabular settings.

In addition, there have been efforts to hybridize explicit and implicit approaches by developing models that combine geospatial distance functions with powerful ML learners, achieving promising results by leveraging the strengths of both paradigms [13] [14]. Despite the growing availability of modern approaches, however, traditional domains such as biology and agriculture still heavily rely on statistical geospatial models that explicitly learn spatial dependencies.

While these models offer strong interpretability and a theoretically sound framework, they often struggle with balancing predictive accuracy and computational efficiency, especially when dealing with large datasets or nonlinear patterns.

Conversely, ML and TDL are designed specifically for large datasets and non-linear distribution learning, which can offer better predictions more efficiently. Although we consider ML and TDL as competitive alternatives to explicit learning approaches, there are still some concerns regarding the hyperparameter tuning cost and potential overfitting issues of TDL. Moreover, due to the inherently spatially correlated nature of geospatial learning, the effectiveness of ML and TDL, which assume that instances are independent obviously raises questions. Therefore, to assess the feasibility of replacing explicit modeling with implicit learning in geospatial tasks and to investigate the effectiveness and efficiency across different model families, we conduct a comprehensive comparison of geospatial statistical models (e.g., Kriging and GWR), ML models (with a focus on tree ensembles), hybrid kernel-based models, and TDL model, e.g., TabPFN, NODE, GATE and etc.

In summary, this work makes the following key contributions:

- We conduct a comparative study across statistical, ML, hybrid, and TDL models to assess predictive performance and training efficiency, with a particular focus on how coordinates are used as input;
- We analyze the practical implications of training and tuning these models in real-world geospatial applications, using a wide range of datasets;
- We perform an exhaustive analysis that includes statistical significance testing and feature importance interpretation, shedding light on how explicit and implicit models utilize spatial information;
- To the best of our knowledge, this is the first benchmark study to systematically evaluate the use of geographic inputs across a broad range of models and datasets;
- All source code and datasets used in this study are publicly available on our GitHub page [15].

This paper is an extended version of our previous publication [1], presented at GEOProcessing 2025. The overall structure is as follows: Section II introduces the relevant methodologies used in geospatial modeling. Section III outlines the experimental setup, including datasets, models, hyperparameter configurations, evaluation metrics and significant test. Section IV presents and discusses the results, highlighting feature contributions in each model. Section V concludes the paper and outlines directions for future work.

II. Methodology Review

Prior benchmark studies have evaluated the performance of various models, ranging from classical ML approaches [16] to Deep Learning (DL) models [17] [18], across both real-world and synthetic datasets [19]. The most recent advancement at the time of writing, TabArena [20], has moreover introduced a dynamic benchmarking platform for tabular data that continuously integrates newly released datasets and models.

While these benchmarks provide exhaustive comparisons of model performance on tabular data, few studies examine how different types of specific feature inputs, particularly spatial features in this case, impact learning. In general-purpose tabular learning, this may not be a critical concern. However, in geospatial learning, where location information plays a central role in capturing spatial autocorrelation, the way spatial features are used becomes highly consequential. To address this gap, we conduct a benchmark study that focuses on the utility of coordinate-based features. Specifically, we compare the effectiveness and efficiency of explicit versus implicit spatial learning paradigms, laying a foundation for understanding their respective advantages in geospatial inference tasks.

This section presents an overview of the underlying mechanisms of the geospatial statistical models, machine learning approaches, hybrid models, and TDL methods evaluated in our work, with an emphasis on their distinct strategies for modeling spatial dependence.

A. Spatial Dependence-Based Models

Kriging and GWR are the most representative models in this group. Although they both heavily rely on the principle of spatial dependence, where observations close to each other are considered more similar than those farther apart, the emphasis of spatial relationships modeling of these two models is slightly varied.

1) Kriging: The main goal of Kriging is to quantify **spatial autocorrelation** to model and estimate the target values by using a variogram, based on the assumption of a jointly Gaussian distribution of the data, followed by computing optimal weights for predictions by solving a system of linear equations, generating the linear unbiased estimates.

The Kriging [21] predictor can be defined as:

$$\hat{Z}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i),$$

where:

- $Z(\mathbf{s}_i)$: observed value at location $\mathbf{s}_i$,
- λ_i: weight assigned to $Z(\mathbf{s}_i)$, determined by spatial correlation.
- n: number of observed locations.

The spatial correlation between locations is modeled using a **variogram** [22], which is defined as:

$$\gamma(h) = \frac{1}{2}\text{Var}[Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h})],$$

where:

- h: distance between two locations,
- $\gamma(h)$: semi-variance at lag h.

By using the variogram, we can calculate the covariance matrix to solve the Kriging system:

$$C(\mathbf{s}_i, \mathbf{s}_j)\Lambda = C(\mathbf{s}_i, \mathbf{s}_0),$$

where Λ indicates the weight assigned to known nodes for the interpolation of an unknown node $\mathbf{s}_0$.

Based on the definition above, Kriging provides an estimate of prediction uncertainty that is defined as:

$$\sigma^2_{\text{Kriging}}(\mathbf{s}_0) = C(\mathbf{s}_0, \mathbf{s}_0) - \sum_{i=1}^{n} \lambda_i C(\mathbf{s}_i, \mathbf{s}_0) - \mu.$$

2) GWR: Compared with Kriging, which focuses on spatial autocorrelation and estimation of the proximity similarity, GWR [23] is based on the assumption of spatial heterogeneity. Though GWR also utilizes the distance matrix as weights to model the spatial variation, it fits a separate regression model locally at each location, weighting observations based on their proximity using a kernel function (e.g., Gaussian or bisquare), which allows for spatial variation in relationships between dependent and independent variables.

Essentially, GWR can be defined as a linear combination:

$$y_i = \beta_0(\mathbf{s}_i) + \sum_{k=1}^{p} \beta_k(\mathbf{s}_i) x_{ki} + \epsilon_i,$$

where:

- y_i: dependent variable at location $\mathbf{s}_i$,
- $\beta_0(\mathbf{s}_i)$ and $\beta_k(\mathbf{s}_i)$: intercept and coefficient (for the k-th independent variable) at location $\mathbf{s}_i$,
- x_{ki}: independent variable at location $\mathbf{s}_i$,
- ϵ_i: random error term at location $\mathbf{s}_i$,
- p: number of independent variables.

The regression coefficients $\boldsymbol{\beta}(\mathbf{s}_i)$ are estimated by solving the weighted least squares problem, which is expressed as

$$\boldsymbol{\beta}(\mathbf{s}_i) = \left(\mathbf{X}^\top \mathbf{W}(\mathbf{s}_i) \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W}(\mathbf{s}_i) \mathbf{y},$$

where $\mathbf{W}(\mathbf{s}_i)$ represents the diagonal weight matrix of the weights assigned to the location, which is close to the point of interest.

To estimate the weight matrix, two kernel functions are commonly used:

- Gaussian kernel:

$$w_{ij} = \exp\left(-\frac{d_{ij}^2}{2b^2}\right),$$

- Bisquare kernel:

$$w_{ij} = \begin{cases} \left(1 - \left(\frac{d_{ij}}{b}\right)^2\right)^2 & \text{if } d_{ij} \le b, \\ 0 & \text{if } d_{ij} > b, \end{cases}$$

where:

- d_{ij}: distance between locations $\mathbf{s}_i$ and $\mathbf{s}_j$,
- b: bandwidth parameter controlling the spatial extent of the weights.

Classical GWR models the local geospatial variation under the assumption of the same spatial scale, while a modification of GWR, namely Multiscale Geographically Weighted Regression (MGWR) [24], provides a more flexible framework by allowing different processes to operate at different spatial scales.

Although Kriging and GWR are widely used for spatial inference tasks, the application scenarios are slightly different. Kriging is typically applied in spatial interpolation, such as estimating soil properties [25], pollutant concentrations [26] [27], or precipitation levels [28], while GWR is commonly applied in spatial regression scenarios, such as modeling house prices [29], predicting socioeconomic factors [30], or environmental influences [31], where relationships vary spatially.

B. Machine Learning Models

Machine learning methods provide a data-driven approach to modeling, focusing on capturing patterns and relationships within the data without explicit assumptions about spatial dependence.

Typically, given a dataset $\{X, Y\}$ consisting of instances $\{x_i, y_i\}$ from a certain distribution $P(Y|X)$, the goal is to learn a function f that maps input features $\mathbf{x} \in \mathbb{R}^d$ to an output $y \in \mathbb{R}$. The general objective is:

$$\hat{f} = \arg\min_f \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i)),$$

where:

- $\ell(y_i, f(\mathbf{x}_i))$: loss function measuring the error between predicted $f(\mathbf{x}_i)$ and actual y_i,
- n: number of training instances.

To minimize the loss function (e.g., mean squared error for regression or cross-entropy for classification), a wide range of optimization algorithms, such as gradient descent and tree-based heuristics were developed to capture complex linear or nonlinear relationships between features. Specifically, tree ensemble models often outperform simpler models on structured data by building a series of decision trees iteratively to minimize the overall loss,

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \gamma_m h_m(\mathbf{x}),$$

where:

- $f_m(\mathbf{x})$: prediction at iteration m,
- $h_m(\mathbf{x})$: weak learner (e.g., a shallow decision tree),
- γ_m: Step size for the weak learner.

Unlike the spatial dependence-based models, which integrate the geospatial information explicitly, ML models are available for all kinds of tabular data inference tasks, but can

be easily applied to the geospatial field by simply including location information (i.e., coordinates in most cases) as features, potentially with further feature-engineering efforts (e.g., distances to landmarks, elevation, land use types, aggregated census information, etc.).

C. Hybrid Kernel-Based Models

Recent advances have sought to explore hybrid approaches to boost the strengths of handling of spatial dependence. The most straightforward trail is to consider Kriging as an extension of GWR, but train these two components separately. Following this basic hybrid idea, Geographically Weighted Regression Kriging (GWRK) [32] was developed and its efficiency proven on datasets from different domains [33] [34].

Another possible combination is to merge Kriging with ML models. By using Kriging as the base model and ML models as either internal learners for residuals [35] or as a super learner [13], this hybrid approach helps mitigate the limitations of both model types, allowing effectively incorporating spatial relationships while enhancing predictive performance.

Moreover, the variogram function in Kriging or a local linear function are not the only choices to model geospatial dependence. GPs can also model spatial dependencies explicitly through kernel functions and by weighting proximal observations spatially. The Gaussian kernel is defined as:

$$k(\mathbf{s}_i, \mathbf{s}_j) = \exp\left(-\frac{\|\mathbf{s}_i - \mathbf{s}_j\|^2}{2\ell^2}\right),$$

where:

- $k(\mathbf{s}_i, \mathbf{s}_j)$: covariance between points $\mathbf{s}_i$ and $\mathbf{s}_j$,
- ℓ: length scale parameter, determining how quickly the correlation decays with distance,
- $\|\mathbf{s}_i - \mathbf{s}_j\|$: Euclidean distance between points $\mathbf{s}_i$ and $\mathbf{s}_j$.

In theory, by embedding spatial correlation into ML workflows, these kernel-based methods enhance predictive performance while retaining the capacity to model non-linearities and complex interactions.

D. Tabular Deep Learning

Similar to traditional ML, TDL specifically targets tabular data using neural networks trained via backpropagation. A diverse range of model architectures have been developed to effectively handle the heterogeneous nature of tabular features, including Bayesian Neural Networks for uncertainty modeling, transformer-based architectures and attention mechanisms for capturing complex feature interactions, and ensemble-gated models that aim to combine the strengths of tree-based ensembles with neural networks through learned feature selection gates. We provide more detailed discussion on these approaches in the following sections.

1) TabPFN: TabPFN is a Transformer-based model that is pre-trained to carry out probabilistic inference under a carefully designed Bayesian-neural-network prior. Building on the PFN framework [36], it can draw direct samples from, and thus closely approximate, the posterior predictive distribution. Unlike classical neural networks or tree ensembles, whose expressiveness is limited by fixed layer counts or tree depth, TabPFN [10] enriches its prior with ideas from Bayesian neural networks [37], [38] and structural causal models [39], [40]. This combination lets it capture intricate feature dependencies and reason about causal relationships in tabular data. It's authors performed a thorough empirical study and showed that TabPFN delivers state-of-the-art inference accuracy. As a pre-trained Transformer, at inference time, TabPFN tokenises every input feature, including coordinates, and processes these tokens through the Transformer's feed-forward layers, yielding calibrated, sample-efficient predictions.

2) NODE: NODE [11] are inspired by gradient-boosted decision trees and use Oblivious Decision Trees (ODTs) as the fundamental building block. In ODTs, all nodes at a given depth split on the same feature and threshold, allowing for a structured and differentiable representation. Each ODT in the NODE architecture splits features based on a shared threshold at a given depth d, and the model outputs the sum of scaled leaf responses. Unlike traditional tree ensembles, NODE requires differentiability. Therefore, the discrete feature-splitting mechanism is replaced by a continuous variant, defined as:

$$\hat{f_i}(x) = \sum_{j=1}^{n} x_j \cdot entmax_\alpha(F_{ij}),$$

where:

- F: feature selection matrix,
- $entmax_\alpha$: α-entmax transformation [41].

The output of each layer is composed as a concatenation of the outputs of m individual trees $\left[\hat{h_1}, \hat{h_2}, \hat{h_3}, \ldots \hat{h_x}\right]$.

3) GATE: GATE [12] is a neural tree-based model that enhances the interpretability and performance of decision trees by incorporating neural network techniques such as gating and residual learning. Like NODE, it employs ODTs to learn nonlinear functions, but instead of averaging outputs, GATE uses Gated Feature Learning Units (GFLUs) to dynamically select and weight feature contributions.

The model output is defined as:

$$\hat{y} = \sigma\left(\sum_{i=1}^{M} \eta_i Y_i\right),$$

where:

- $\hat{y}$: final output,
- η: learnable parameters from gate mechanism,
- Y: the output of each sub-trees.

4) FT-Transformer: FT-Transformer [42] is another influential Transformer based model, tailored for tabular data. It combines a feature tokeniser with a transformer encoder, converting both categorical and numerical features into dense embeddings before applying multiple Transformer layers.

Given an input x, each feature x_j is first transformed to an embedding:

$$T_j = b_j + f_j(x_j) \in \mathbb{R}^d \qquad f_j : \mathbb{X}_j \to \mathbb{R}^d$$

where:

- b_j: bias term,
- f_j: mapping function to map the feature to an embedding space.

After concatenating all embeddings for both categorical and numerical features, a series of Transformer layers F_i are applied sequentially:

$$T_i = F_i(T_{i-1}),$$

producing the final feature representation used for prediction.

5) DANets: DANets [43] build upon the FT-Transformer architecture by introducing instance-based attention. While maintaining the transformer backbone, DANets extend the attention mechanism by incorporating a learnable sparse mask in their abstract layer, enabling the model to capture complex and hierarchical feature interactions.

Given an input vector $f \in \mathbb{R}^m$ where m indicates feature numbers, a learnable sparse mask $M \in \mathbb{R}^m$ is used to filter the features by by element-wise multiplying with the vector f. The feature selection is defined as,

$$M = entmax_\alpha(W_{mask}), \qquad f' = M \odot f$$

where:

- $entmax_\alpha$: α-entmax transformation (same as NODE),
- W_{mask}: learnable vectors,
- M: a learnable sparse mask.

In conclusion, all models discussed above leverage geospatial dependence by either integrating spatial relationships through distance matrices, modeling interactions between a target point and its neighbors, or engineering spatial proximity as explicit input features, without directly modeling spatial autocorrelation.

While numerous real-world applications adopt these methods, the comparative efficiency and benefits of explicitly modeling spatial dependence remain underexplored. Traditional models such as Kriging and GWR rely heavily on spatial distance matrices, often leading to high computational costs and numerical issues, such as singular matrices, which makes them less scalable. In contrast, ML and TDL approaches avoid solving systems of equations derived from spatial relationships. Instead, they learn a direct mapping from feature space to the target variable, efficiently handling large datasets. However, they come with their own challenges, such as overfitting risks and increased computational burden from backpropagation.

To systematically assess the trade-offs, we conduct an extensive benchmark study evaluating each model's predictive performance and computational efficiency on geospatial inference tasks. Additionally, we apply feature attribution techniques to quantify the contribution of spatial features (e.g., coordinates), providing deeper insights into their role in model performance. We hope this study helps inform practical model selection for real-world geospatial applications, especially in light of increasingly powerful ML and TDL models.

III. Experimental Setup

In this section, we describe an exhaustive experiment to compare a wide range of ML models, statistics models, and TDL models, covering a collection of real-life datasets, with a complementary explanation of comparison hyperparameters, significance test and the interpretation approach.

A. Datasets

We utilize a vast collection of public datasets, primarily comprising real estate valuation datasets sourced from Kaggle, biology-related datasets from the R package `Spatstat.data`, and one additional well-known "yield" dataset [44].

TABLE I
Dataset summary with number of instances and features

Dataset	Nr. Instances	Nr. Features
singapore	9212	6
london	34994	10
melbourne	5759	12
newyork	4170	6
paris	21765	6
beijing	3745	13
perth	30210	9
seattle	20832	15
dubai	406	9
yield	1696	24
anemones	231	2
bronzefilter	678	2
longleaf	584	2
spruces	134	2
waka	504	2

A summary of the datasets used is provided in Table I, including the number of instances and features. The original features include both numerical and categorical variables, though all categorical features were transformed into a numerical format using the CatBoost encoder.

All datasets include spatial coordinates (either geographic or geometric) along with auxiliary features, such as hedonic attributes in the case of real estate valuation datasets. To investigate the impact of different spatial modeling strategies under different circumstances, we leverage each dataset to construct two variants: (i) coordinates-only, and (ii) all-features (original). Models are then trained on both variants to assess the effectiveness of explicit versus implicit spatial dependence modeling.

Prior to training, we first clean all datasets by converting all categorical features to encoded numerical ones (see remark above). We then continue by removing duplicates and rescaling features to the [0, 1] range to assure an equal playing ground. For those datasets containing timestamps, we then apply a temporal split (i.e., chronological partitioning) to construct

training, validation, and test sets. For datasets without any temporal information, we perform a random split strategy. In either cases, we allocate 70% of the data for training, 10% for validation to be used for hyperparameter tuning, and 20% for testing and establish final performance metrics.

To ensure reliable geospatial inference, we carefully preprocess spatial coordinates as follows. All coordinates are converted into a Cartesian coordinate system, tailored to the geographic location of each dataset. This transformation standardizes the spatial input for all models and avoids distortions that may occur with spherical geometries, which is particularly relevant for statistical models relying on distance matrices. Specifically, for GWR and Kriging, these Cartesian coordinates remain unscaled to preserve accurate Euclidean distance computations whenever this distance metric is used, whilst for ML and TDL models, the coordinates are scaled in the same way as the other features. For the sake of notation, the unscaled Cartesian coordinates will be denoted as "lat" and "lon", whereas the scaled ones will be labeled as "x" and "y" henceforth.

B. Models

We select a broad set of models for inclusion in our comparative study, spanning traditional ML, TDL, kernel-based methods, and geospatial statistical models. Serving as *statistical baselines*, Kriging [45] [46] [47] and GWR [23] are naturally included.

Alongside, we evaluate *machine learning* models, including the representative Linear Regression of Ridge regularization [48], and Support Vector Machines (SVM) [49] [50] [51]. In particular, we place special emphasis on tree ensemble methods, including Random Forest (RF) [52], XGBoost [53], LightGBM (LGBM) [54], and CatBoost [55]. These models have demonstrated consistently strong performance in practice, and in many benchmark studies on tabular data, they have even outperformed deep learning approaches [56].

To assess the benefit of *hybrid kernel-based architectures*, we include Kriging-LGBM [35], a two-stage model that uses a LightGBM regressor as the primary learner and applies Kriging to fit the residuals. Additionally, we include Gaussian and Power Tweedie. These two models are based on parametric assumptions about the target distribution (e.g., normal, Poisson-Gamma), while GPs explicitly incorporate spatial correlation using kernel functions that capture similarity between coordinate-based inputs.

The final group focuses on *TDL models*. As an extension of our previous work (where we focused solely on TabPFN), this study expands the scope to a much broader range of TDL architectures, including FT-Transformer, DANet, Gated Adaptive Network for Deep Automated Learning of Features (GANDALF) [57], GATE, and NODE.

To ensure a fair comparison across all models, we conduct an exhaustive hyperparameter tuning process. The hyperparameter grid used for all models is detailed in Table II. It is worth noting that not all models support or require extensive tuning. For instance, TabPFN, being a pre-trained model, is designed to achieve competitive results out-of-the-box by leveraging pre-trained weights without additional tuning. Moreover, other TDL models utilize an automatic learning rate optimization mechanism by default and are hence not listed separately.

C. Evaluation

All models are evaluated from two complementary perspectives: (i) predictive performance and (ii) computational cost (i.e., training time).

For predictive performance, we adopt Root Mean Square Error (RMSE) as the primary evaluation metric. Each model is initially trained on the training set, followed by systematic hyperparameter tuning on the validation set using RMSE as the selection criterion. The best-performing configuration is then used to evaluate the resulting final model's predictive performance on the unseen test set. All models share identical data partitions and are assessed using the same consistent evaluation procedure.

To further investigate the role of spatial information in predictive performance, we evaluate each model under two distinct data configurations, as mentioned above, either using only spatial coordinates as features ("Coordinates-only"), or incorporating both coordinates and all additional attributes, when available ("All-features"). This comparison aims to highlight the importance of spatial information and to evaluate the models' ability to capture spatial dependence, either explicitly or implicitly.

In terms of computational cost, we record the total training time required for each model across the entire hyperparameter tuning process, along with the number of tuning rounds. We then calculate the average training time per tuning round as the final comparison metric. This analysis provides insights into the computational efficiency of each method, enabling a comprehensive assessment of the trade-off between model accuracy and training overhead.

All the experiments are conducted on a standard workstation equipped with an Intel Core i9-13900 (13th Gen) CPU, 64 GB RAM, and an NVIDIA RTX A5000 GPU. All traditional machine learning models are trained using CPU resources, while the TDL models are trained with GPU acceleration.

D. Statistical Testing

To compare the performance of multiple algorithms across multiple datasets, we adopt the widely used Demšar method [58]. This approach is designed to assess whether the differences in performance among models are statistically significant by computing average ranks and visualizing them with a Critical Difference (CD) diagram.

Compared to parametric alternatives such as the paired t-test or ANOVA, the Demšar analysis is non-parametric, making no assumptions about data distribution and being especially suited for small-sample evaluations and providing robustness from its reliance on rank ordering rather than raw values. The method starts from ranking each model's performance on every dataset, and then calculating average

TABLE II
OVERVIEW OF MODELS AND THEIR HYPERPARAMETERS USED IN THE COMPARISON.

Category	Type	Model	Hyperparameters
Geospatial Statistics	Geospatial Heterogeneity	GWR	best bandwidth for kernel
	Geospatial Autocorrelation	Kriging	nlags: [30, 60, 90, 120]
			variogram_model: [“gaussian”, “linear”]
Machine Learning	Linear	Ridge LR	α: [0.1, 0.2,. . . , 0.9]
		SVM	C: [1, 11,. . . , 101]
			ϵ: [0.1, 0.2,. . . , 0.9]
	Tree Ensemble	RandomForest	min_samples_split: [2, 3, 5]
			min_samples_leaf: [3, 5, 10]
		XGBoost	learning_rate: [0.1, 0.01, 0.005]
			reg_alpha: [0.0, 0.1,. . . , 1.0]
			reg_lambda: [0.0, 0.1,. . . , 1.0]
		LGBM	learning_rate: [0.1, 0.01, 0.005]
			reg_alpha: [0.0, 0.1,. . . , 1.0]
			reg_lambda: [0.0, 0.1,. . . , 1.0]
		CatBoost	iterations: [100, 200]
			learning_rate: [0.001, 0.005, 0.01, 0.05, 0.1]
			l2_leaf_reg: [0.1, 0.5, 1, 5]
Kernel-Based	Gaussian	Gaussian Process	kernel: C(1.0) * RBF(length_scale_bounds=(1e-2, 1e2))
			alpha: [0.1, 0.2,. . . , 0.9]
	Power	Tweedie	power: [0, 1, 1.2, 1.5, 1.8, 2, 3]
			alpha: [0.0, 0.1,. . . , 0.9] + [2, 5, 8, 10]
	ML Kernel	Kriging LGBM	Kriging parameters (same as base Kriging):
			nlags = [30, 60, 90, 120]
			variogram_model: [“gaussian”, “linear”]
			LGBM parameters:
			reg_alpha: [0.0, 0.5, 1.0]
			reg_lambda: [0.0, 0.5, 1.0]
			learning_rate: [0.1, 0.01, 0.005]
Deep Learning	Tabular DL	TabPFN	—
		FT-Transformer	num_heads: [4, 8, 16]
			attn_dropout: [0.0, 0.1, 0.2, 0.4]
		DANet	n_layers: [8, 20]; k: [3, 5, 8]
			dropout_rate: [0.1, 0.2, 0.3]
		GANDALF	gflu_stages: [2, 4, 6, 8, 10]
			gflu_dropout: [0, 0.1, 0.2, 0.3]
		GATE	gflu_stages: [2, 4, 6, 8, 10]
			gflu_dropout: [0, 0.1, 0.2, 0.3]
		NODE	num_layers: [1, 2, 4]
			num_trees: [8, 16, 32, 64]
			depth: [3, 4, 6]
			input_dropout: [0, 0.1, 0.2, 0.3]

ranks across all datasets. Next, a Friedman test is applied to assess whether there are any overall differences between the models. If such significance is detected, a post-hoc Nemenyi test is conducted to compare all pairs of models. Two models are considered significantly different if the difference in their average ranks exceeds a computed CD:

$$CD = q_\alpha \cdot \sqrt{\frac{m(m+1)}{6n}}$$

where:

- q_α: critical value from the Studentized range distribution in Nemenyi test,
- m: number of all models,
- n: number of all datasets,
- α: significance level.

We employ this framework to visualize outperforming models, with the detailed results presented in Section IV below.

E. Interpretation

To further investigate the role of individual features, particularly geographical coordinates, in geospatial learning, we conducted an interpretation analysis using SHAP (SHapley Additive exPlanations) values [59]. SHAP, grounded in game theory, provides a consistent method to quantify how much each feature contributes to a model's prediction by treating each feature as a player in a cooperative game and computing its marginal contribution across all feature subsets, which can be simply defined as a linear combination,

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i$$

where:

- $f(x)$: the prediction function,
- ϕ_0: average predictions of the model,
- M: number of all features,
- ϕ_i: shap values of each feature i.

And the Shapley values ϕ_i is defined as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} (\hat{f}(S \cup \{i\}) - \hat{f}(S))$$

where:

- S: all possible subsets with all features except the i^{th} one,
- M: number of all features,
- $\hat{f}(S)$: predictions on subset.

IV. Results

In this section, we present our experimental results, encompassing both predictive performance and computational cost across multiple geospatial datasets. Additionally, we include a statistical model ranking analysis. A dedicated subsection is also provided for interpretability, using SHAP values to explore the contributions of different features to model predictions.

A. Performance & Cost

1) Performance: Table III reports the results across all datasets under the two configurations of the datasets: (i) coordinates and additional features (denoted as "Dataset (All)"), and (ii) only spatial coordinates (denoted as "Dataset (Coord)").

Among all models, TabPFN consistently achieves the lowest RMSE across both data configurations. Particularly for datasets with additional features, TabPFN outperforms all other models in nearly all cases, highlighting its strong generalization ability and effectiveness in handling geospatial inference tasks.

By comparison, other TDL architectures, namely FT-Transformer, DANet, GANDALF, and GATE, fall short of TabPFN's accuracy. Only NODE shows competitive behavior, outperforming the baseline on two coordinate-only settings. These findings imply that TDL models are not inherently superior, especially on the small, data-sparse problems typical of geospatial analysis. At the same time, TabPFN's recency highlights the considerable headroom that still exists for TDL methods in this setting.

Meanwhile, tree ensemble models, notably LightGBM, RMF, and CatBoost, exhibit consistently strong performance, frequently ranking first or second. Their competitive performance, even compared to TDL models, underscores the robustness of ensemble-based learners for tabular geospatial data. Indeed, this finding is in line with general prior studies [56] that conform the strength of these models, but highlights in our context that they are well capable to treat coordinates as any other feature.

In contrast, linear models, such as Ridge Regression, generally perform worse than statistical geospatial models like Kriging and GWR. This further confirms the limitation of simple linear assumptions in order to capture spatial heterogeneity.

The statistical models, GWR and Kriging, however, do not outperform TabPFN or ensemble models overall but do demonstrate notable effectiveness in the coordinates-only setting. This aligns with their theoretical strengths in capturing explicit spatial autocorrelation, particularly when supplementary features are unavailable.

In summary, TabPFN offers state-of-the-art performance across both feature configurations, proving its capacity to implicitly learn spatial dependencies. This finding is in line with our previous work [1], but our exhaustive study performed here shows that other TDL approaches do not achieve the same result. Next, tree ensemble models remain highly effective and computationally efficient alternatives, especially when training data is limited. Finally, statistical models (GWR and Kriging) do retain their relevance in coordinate-only scenarios, validating their role in explicitly modeling spatial structure.

More importantly, it is notable that the comparison between datasets with and without additional features confirms the importance of complementary information. We see that across

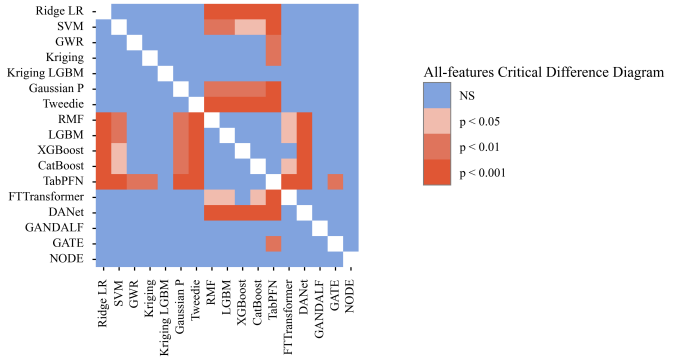


Figure 1. All features: visualizations of Demšar analysis on performance significance.

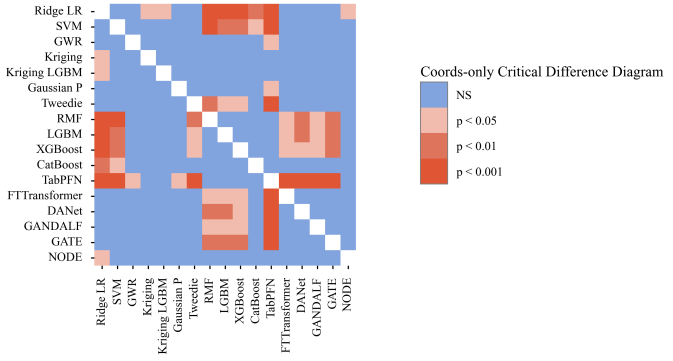


Figure 2. Coordinate features: visualizations of Demšar analysis on performance significance.

nearly all models, the inclusion of additional attributes leads to substantially lower RMSE, emphasizing the necessity of domain-specific features being available, and the benefit of rich feature engineering in geospatial learning.

2) Significance: To support our comparison results, we conduct a statistical significance analysis using the method proposed by Demšar [58]. First, the average rank of each model was computed across all datasets. After confirming the significance of a Friedman test, We then applied a Nemenyi's post-hoc test for pairwise comparisons, identifying statistically significant differences between models based on whether the difference in their average ranks exceeded the CD threshold.

The results are visualized in Figure 1 and Figure 2, where the X and Y-axes denote the model names. Each grid cell is color-coded to reflect the level of statistical significance between each model pair.

When using datasets with both coordinates and additional features, TabPFN significantly outperforms all other models, including both statistical models (GWR and Kriging) and other deep learning techniques. Whereas tree ensemble models (e.g., LightGBM, Random Forest, and CatBoost) do not achieve statistically significant improvements over GWR or Kriging, they do significantly outperform weaker models such as Gaussian Process and Tweedie Regressor, both of which explicitly encode spatial distance in their modeling.

For datasets with only spatial coordinates, fewer statistically significant differences are observed across models. Even the best-performing model, TabPFN, is not significantly better than Kriging, though it still significantly outperforms GWR. This result reinforces the value of explicit spatial modeling techniques (e.g., Kriging) when no additional features are available.

We also present two CD diagrams in Figure 3 (for datasets with all features) and Figure 4 (for coordinates-only datasets), which illustrate the average rank of each model across all datasets. As a refresher: models are ordered from best (left) to worst (right) along the X-axis. Each model is marked with a star to indicate its rank. Horizontal bars are used to group models that do not differ significantly from each other.

In the all-feature setting, TabPFN achieves the highest average rank and is significantly better than statistical models. The family of tree ensemble models follow closely and, although not statistically superior to TabPFN, are substantially better than linear models, most TDL models and statistical models, which occupy the lowest ranks.

In the coordinates-only setting, TabPFN again ranks first but is not significantly better than Kriging, corroborating earlier findings that explicit spatial models remain competitive when limited to spatial coordinates alone.

3) Computational Cost: We assess computational cost in terms of training time, as shown in Figure 5 and Figure 6. Training efficiency is a critical factor for the deployment of geospatial models in practice, especially in large-scale or resource-constrained environments.

Traditional geospatial statistical models, such as GPs, Kriging, and GWR, incur significantly higher computational costs, with training time increasing exponentially as dataset size grows. These models consistently record the longest training durations across all experiments, primarily due to their reliance on spatial dependence structures and costly matrix operations.

In contrast, TabPFN achieves high efficiency owing to its pre-trained foundation model, requiring reasonable training and tuning time when used on a new dataset. This, combined with its superior predictive accuracy, positions TabPFN as a highly effective model that balances both performance and efficiency.

While other TDL models (e.g., transformer-based or attention-based architectures) benefit from GPU acceleration, they still exhibit considerably higher training costs compared to traditional machine learning models. Despite this, TDL models generally outperform statistical baselines when additional features are included.

Among traditional machine learning models, tree ensemble methods (e.g., LightGBM, Random Forest, CatBoost) demonstrate a favorable trade-off between computational cost and predictive accuracy. These models benefit from optimized implementations and efficient tree construction algorithms, leading to relatively low training times even on larger datasets.

Overall, our results suggest that explicit spatial modeling is not always necessary, particularly when strong predictive

TABLE III
COMPARISON OF MODEL PERFORMANCE (RMSE) ACROSS DIFFERENT DATASETS. RESULTS ARE GROUPED BY WHETHER ADDITIONAL FEATURES ARE USED OR ONLY COORDINATES. BEST SCORES ARE IN **BOLD**, SECOND-BEST IN UNDERLINE.

Dataset (All)	Ridge LR	SVM	GWR	Kriging	Kriging LGBM	Gaussian P	Tweedie	RMF	LGBM	XGBoost	CatBoost	TabPFN	FTTransformer	DANet	GANDALF	GATE	NODE
beijing	0.1718	0.1393	0.1335	0.1284	0.1285	0.1608	0.1693	0.1034	0.1003	0.1045	0.1050	0.1009	0.1403	0.3104	**0.0959**	0.1659	0.1000
dubai	0.1801	0.1981	0.1852	0.1373	0.1303	0.1982	0.1905	0.1218	0.1202	0.1201	0.1112	**0.1042**	0.1776	0.2127	0.1850	0.1899	0.1377
london	0.0824	0.0717	0.0659	0.0637	0.0621	0.0732	0.0824	0.0577	0.0574	0.0583	0.0595	**0.0557**	0.0666	0.0827	0.0655	0.0646	0.0622
melbourne	0.0803	0.0702	0.0673	0.0603	0.0389	0.0687	0.0581	0.0327	0.0296	0.0313	0.0289	**0.0263**	0.0592	0.0944	0.0472	0.0435	0.0391
newyork	0.0868	0.0742	0.0858	0.0816	0.0684	0.0726	0.0883	0.0575	0.0578	0.0581	0.0584	**0.0567**	0.0736	0.0797	0.0694	0.0823	0.0634
paris	0.0213	0.0246	0.0206	0.0213	0.0213	0.0217	0.0214	**0.0201**	0.0202	0.0203	0.0202	**0.0201**	0.0216	0.0625	0.0207	0.0214	0.0207
perth	0.0494	0.0459	0.0355	0.0348	0.0324	0.0375	0.0489	**0.0269**	0.0277	0.0275	0.0279	0.0274	0.0383	0.0440	0.0320	0.0337	0.0306
seattle	0.1252	0.1100	0.0967	0.1101	0.0981	0.1134	0.1253	0.0833	0.0820	0.0830	0.0834	**0.0791**	0.0918	0.1034	0.0894	0.0871	0.0872
singapore	0.1057	0.0715	0.0654	0.1090	0.0780	0.0803	0.0975	0.0574	0.0563	0.0575	0.0547	**0.0510**	0.0749	0.0834	0.0722	0.0744	0.0677
yield	0.1314	0.0778	0.1141	0.0554	0.0580	0.0967	0.1278	0.0577	0.0566	0.0550	0.0541	**0.0498**	0.1541	0.1137	0.1349	0.1288	0.1198
Dataset (Coord)	**Ridge LR**	**SVM**	**GWR**	**Kriging**	**Kriging LGBM**	**Gaussian P**	**Tweedie**	**RMF**	**LGBM**	**XGBoost**	**CatBoost**	**TabPFN**	**FTTransformer**	**DANet**	**GANDALF**	**GATE**	**NODE**
beijing	0.1833	0.1342	0.1377	0.1284	0.1284	0.1380	0.1833	0.1294	0.1279	0.1279	0.1279	0.1272	0.1318	0.1304	0.1341	0.1815	**0.1261**
dubai	0.1941	0.1583	0.1692	**0.1373**	**0.1373**	0.1539	0.1911	0.1376	0.1448	0.1413	0.1405	0.1390	0.1701	0.1896	0.1859	0.1935	0.1580
london	0.0858	0.0702	0.0656	0.0637	0.0637	0.0690	0.0858	**0.0627**	0.0640	0.0643	0.0657	0.0641	0.0708	0.0685	0.0699	0.0843	0.0670
melbourne	0.0944	0.0708	0.0758	0.0602	0.0603	0.0652	0.0922	0.0605	0.0610	0.0599	0.0604	**0.0588**	0.0664	0.0945	0.0786	0.0812	0.0621
newyork	0.1060	0.0973	0.0921	0.0908	0.0908	0.0919	0.1062	0.0879	0.0882	0.0886	0.0890	**0.0876**	0.0961	0.0914	0.0928	0.1080	0.0916
paris	0.0216	0.0614	0.0205	0.0213	0.0213	0.0217	0.0216	0.0205	0.0203	0.0203	0.0203	**0.0202**	0.0215	0.0215	0.0208	0.0212	0.0203
perth	0.0555	0.0444	0.0349	0.0348	0.0348	0.0384	0.0548	**0.0339**	0.0340	0.0341	0.0343	0.0340	0.0376	0.0379	0.0366	0.0473	0.0356
seattle	0.1448	0.1181	0.1141	0.1101	0.1101	0.1154	0.1448	**0.1089**	0.1096	0.1096	0.1102	0.1100	0.1190	0.1135	0.1186	0.1236	0.1143
singapore	0.1524	0.1328	0.1334	0.1090	0.1090	0.1401	0.1524	**0.1059**	0.1171	0.1152	0.1207	0.1096	0.1342	0.1361	0.1354	0.1406	0.1333
yield	0.2379	0.0674	0.0805	0.0517	0.0517	0.0732	0.2277	0.0535	0.0549	0.0545	0.0529	**0.0487**	0.1335	0.0876	0.0781	0.2715	0.0685
anemones	0.1756	0.1869	0.1856	0.1826	0.1826	0.1804	0.1755	0.1755	**0.1747**	0.1779	0.1768	0.1808	0.1817	0.2238	0.2126	0.1855	0.1921
bronzefilter	0.1736	0.2364	0.2119	0.1835	0.1835	0.1754	0.1622	0.1555	0.1623	0.1615	0.1814	**0.1535**	0.2849	0.1778	0.1929	0.2635	0.1729
longleaf	0.3114	0.2978	0.2545	0.2750	0.2750	0.2923	0.2531	0.2608	0.2798	0.2639	0.3036	**0.2460**	0.2564	0.2673	0.2573	0.2469	0.2524
spruces	0.2038	0.2361	0.1972	0.2284	0.2284	0.1889	0.1942	0.2167	0.1889	0.2004	0.1911	0.1935	0.1989	0.2167	0.1992	0.1874	**0.1861**
waka	0.1240	0.1398	0.1237	0.1295	0.1295	0.1233	0.1235	0.1282	0.1235	0.1234	0.1237	**0.1232**	0.1319	0.1442	0.1283	0.1236	0.1283

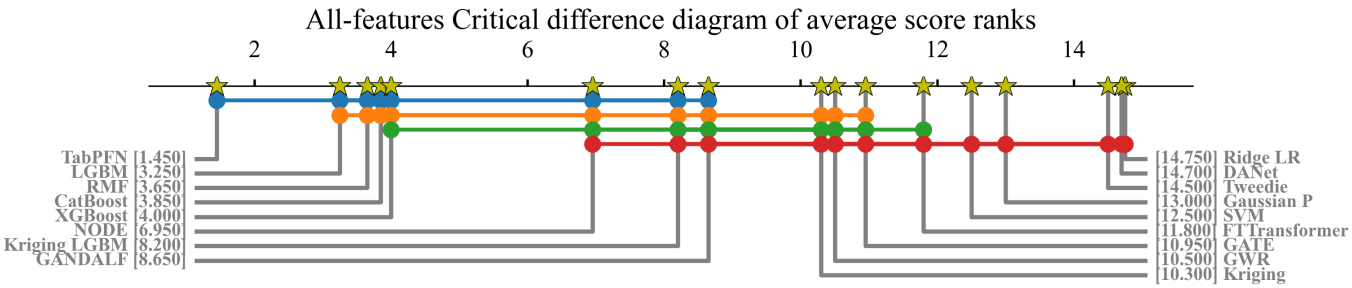


Figure 3. All features: visualizations of Demšar analysis on performance ranking.

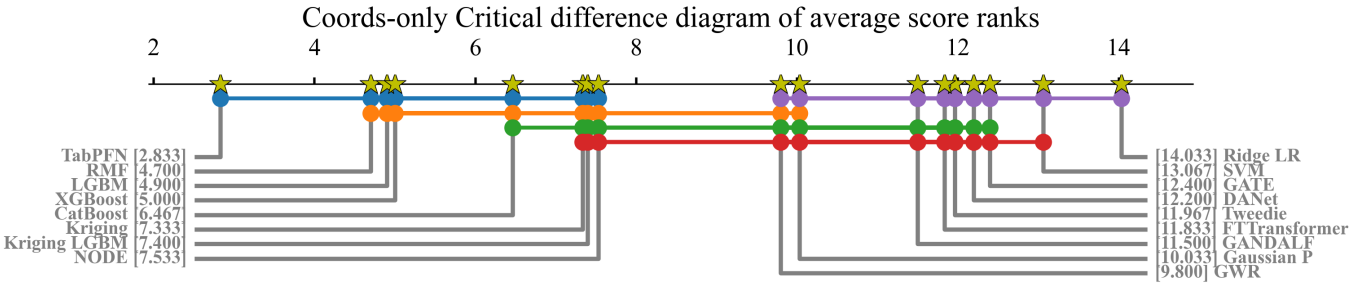


Figure 4. Coordinate features: visualizations of Demšar analysis on performance ranking.

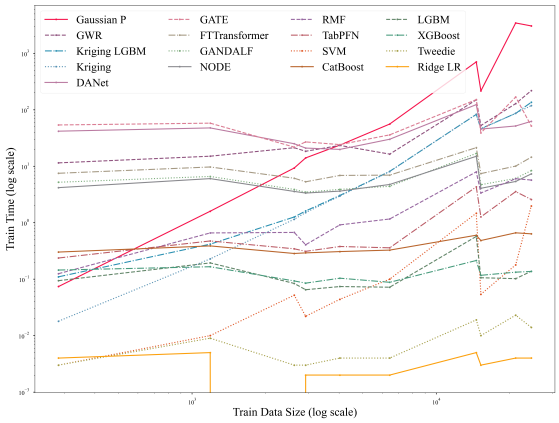


Figure 5. All features: visualizations of training time (s) per hyperparameter run across different models in log scale.

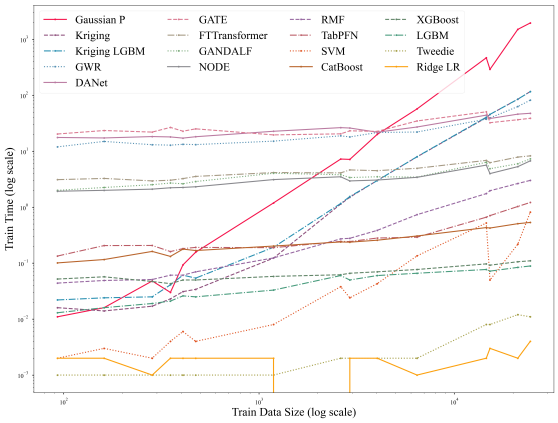


Figure 6. Coordinate features: visualizations of average training time (s) per hyperparameter run across different models in log scale.

models and additional contextual features are available, as is often the case in real-life settings. Tabular learners, including TabPFN and tree ensembles, are capable of implicitly capturing spatial dependencies from coordinate inputs alone. By treating spatial coordinates as standard input features rather than modeling geospatial dependencies explicitly, one can significantly reduce computational overhead while preserving or even improving predictive accuracy. This efficiency advantage is especially valuable for scalable geospatial applications where frequent retraining or rapid deployment is required.

B. Interpretation

Due to the architectural diversity across models, we adopted model-specific SHAP approximations. For deep learning models, we applied Gradient SHAP [59], which is tailored to differentiable models and estimates SHAP values by computing the expectation of gradients relative to randomly sampled baselines. For the other models, including tree ensembles, linear models, and statistical baselines, we employed the Kernel SHAP explainer [59], a model-agnostic method that perturbs inputs and fits a locally weighted linear model to

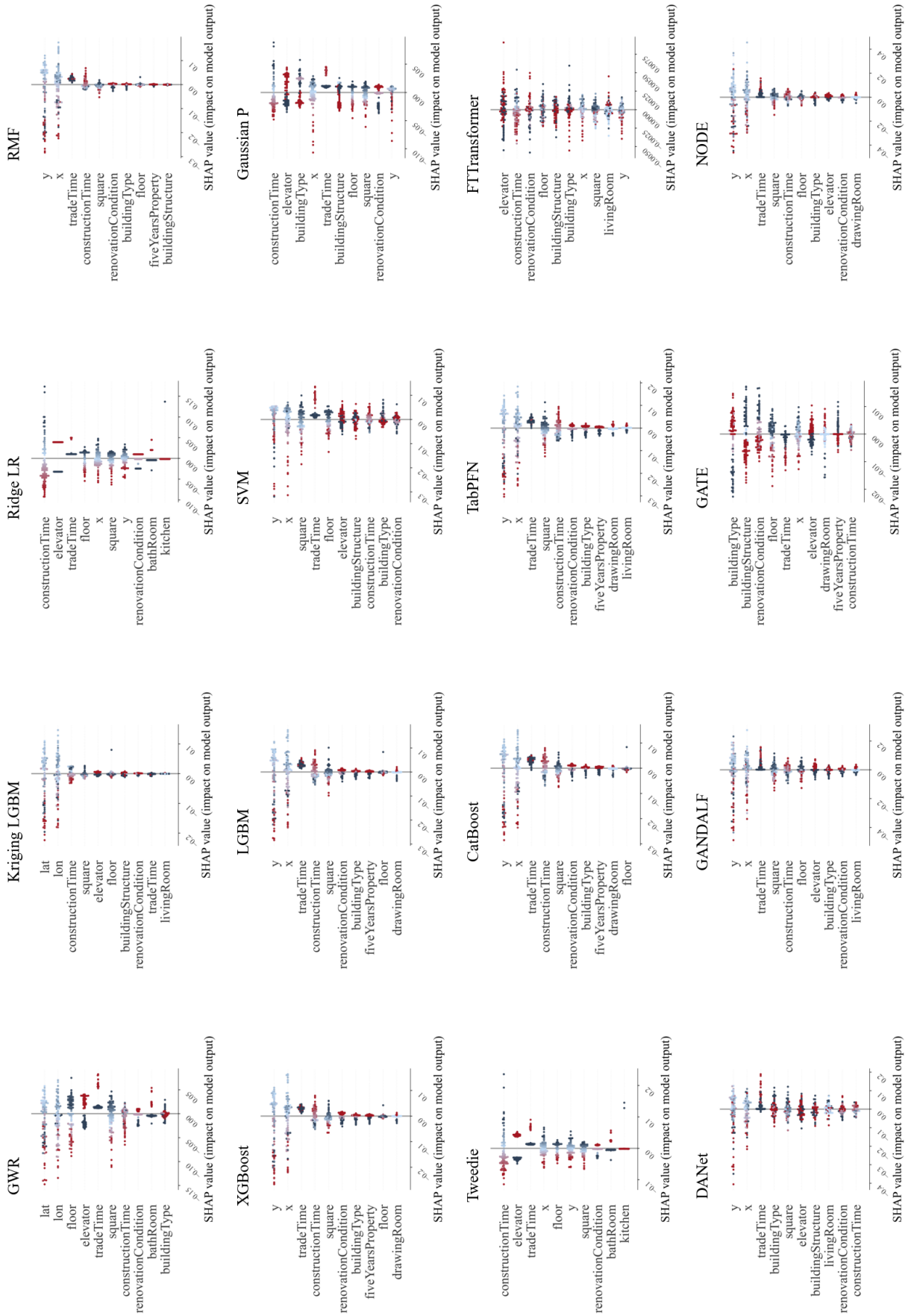


Figure 7. Feature attributions using the Beijing property dataset.

approximate each feature's contribution.

To elaborate this analysis, we select the "Beijing" property dataset and randomly sample 1000 instances as a representative test set. Feature attributions were then computed for all models and ranked based on their importance scores. For models explicitly leveraging geospatial dependencies (e.g., GWR, Kriging), spatial coordinates are labeled as "lat" and "lon". For models treating coordinates as standard tabular features, the coordinate inputs are denoted "x" and "y".

As shown in Figure 7, most models effectively utilize spatial information, either implicitly or explicitly. Noteworthy is that models with weaker predictive performance, such as NODE and Ridge Regression, assign lower importance to coordinate features, indicating their limited ability to extract spatial patterns. In contrast, strong-performing models, such as TabPFN, tree ensembles, and Kriging, consistently rank spatial features amongst the most influential.

These findings reinforce the notion that explicit spatial modeling is not strictly necessary for successful geospatial inference. Modern tabular learners, when sufficiently expressive, can implicitly learn geospatial patterns from coordinate features without specialized geospatial mechanisms. This highlights a promising alternative to traditional geo-statistical modeling, particularly when working with feature-rich tabular datasets.

V. Conclusions

The primary goal of this work was to explore the distinction between explicit and implicit spatial learning in the context of geospatial inference. Traditionally, statistical models such as Kriging and GWR capture spatial autocorrelation explicitly through distance-based functions. However, instead of relying on these computationally intensive distance-based modeling approaches, an alternative is simply to treat coordinates as standard input features and leverage them within ML or TDL models, given that they are strong enough. This implicit strategy can yield competitive predictive performance while significantly reducing computational costs.

To support this argument, we conducted a comprehensive evaluation across a wide range of models for geospatial regression tasks, including traditional statistical approaches, linear and ensemble-based ML methods, and recent TDL architectures. Our benchmark considered both datasets with only coordinate features and those augmented with additional geospatial attributes. The results reveal clear distinctions in predictive accuracy, training efficiency, and interpretability across these different model families.

Our findings indicate that whilst geo-statistical models do remain relevant for specific coordinate-only settings, general-purpose tabular learners, particularly TabPFN and ensemble tree models, emerge as powerful, scalable, and interpretable alternatives for modern geospatial learning. Furthermore, the SHAP-based feature attribution analysis demonstrates that top-performing models such as TabPFN and tree ensembles are capable of leveraging spatial features effectively, and do so without explicit spatial modeling. In contrast, weaker models tend to underutilize spatial information, reinforcing the importance of model capacity in capturing geospatial dependencies.

In summary, our results call for a re-evaluation of the traditional spatial learning paradigm. They demonstrate the feasibility and efficiency of treating location information as standard input features, empowering both ML and TDL models to perform robust geospatial inference in real-world applications.

We strongly extended the scope of models and datasets compared to our previous research. Nevertheless, further work is needed and invited to solidify these findings. Due to the limited availability of large-scale public datasets, we were unable to fully assess TDL models in extreme data-rich settings, where deep learning typically thrives. Additionally, most of our evaluations were conducted in highly urbanized areas, offering limited insight into how these learning paradigms perform in sparsely populated or rural regions. Future research could also explore more hybrid models, or zoom into upcoming transformer-based architectures. Finally, a best-effort fair choice was made in this study in terms of hyperparameter tuning, which could be optimized further across all models to potentially enhance top-performance levels.

References

[1] Y. Guo and S. vanden Broucke, "Coordinates are just features: Rethinking spatial dependence in geospatial modeling," in *GEOProcessing 2025, The Seventeenth International Conference on Advanced Geographic Information Systems, Applications, and Services*, pp. 48–55.

[2] P. K. Rai, V. N. Mishra, and P. Singh, *Geospatial technology for landscape and environmental management: sustainable assessment and planning*. Springer, 2022.

[3] J. K. Thakur, S. K. Singh, A. Ramanathan, M. B. K. Prasad, and W. Gossel, *Geospatial techniques for managing environmental resources*. Springer Science & Business Media, 2012.

[4] B. Jiang and X. Yao, *Geospatial analysis and modelling of urban structure and dynamics*. Springer Science & Business Media, 2010, vol. 99.

[5] L. A. Manfré, E. Hirata, J. B. Silva, E. J. Shinohara, M. A. Giannotti, A. P. C. Larocca, and J. A. Quintanilha, "An analysis of geospatial technologies for risk and natural disaster management," *ISPRS International Journal of Geo-Information*, vol. 1, no. 2, pp. 166–185, 2012.

[6] N. N. Kussul, B. V. Sokolov, Y. I. Zyelyk, V. A. Zelentsov, S. V. Skakun, and A. Y. Shelestov, "Disaster risk assessment based on heterogeneous geospatial information," *Journal of Automation and Information Sciences*, vol. 42, no. 12, 2010.

[7] D. G. Krige, "A statistical approach to some basic mine valuation problems on the witwatersrand," *Journal of the Southern African Institute of Mining and Metallurgy*, vol. 52, no. 6, pp. 119–139, 1951.

[8] G. Matheron, "Principles of geostatistics," *Economic Geology*, vol. 58, no. 8, pp. 1246–1266, 1963.

[9] C. Bitter, G. F. Mulligan, and S. Dall'erba, "Incorporating spatial variation in housing attribute prices: a comparison of geographically weighted regression and the spatial expansion method," *Journal of Geographical Systems*, vol. 9, no. 1, pp. 7–27, Apr. 2007.

[10] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, "Tabpfn: A transformer that solves small tabular classification problems in a second," *arXiv preprint arXiv:2207.01848*, 2022.

[11] S. Popov, S. Morozov, and A. Babenko, "Neural oblivious decision ensembles for deep learning on tabular data," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[12] M. Joseph and H. Raj, "GATE: gated additive tree ensemble for tabular classification and regression," *CoRR*, vol. abs/2207.08548, 2022.

[13] G. Erdogan Erten, M. Yavuz, and C. V. Deutsch, "Combination of machine learning and kriging for spatial estimation of geological attributes," *Natural Resources Research*, vol. 31, no. 1, pp. 191–213, 2022.

[14] Z.-Y. Chen, R. Zhang, T.-H. Zhang, C.-Q. Ou, and Y. Guo, "A kriging-calibrated machine learning method for estimating daily ground-level no2 in mainland china," *Science of The Total Environment*, vol. 690, pp. 556–564, 2019.

[15] ArmonGo, "Github: Geocoordsfeatsext," accessed: March 31, 2025. [Online]. Available: https://github.com/ArmonGo/GeoCoordsFeatsExt

[16] F. Conrad, M. Mälzer, M. Schwarzenberger, H. Wiemer, and S. Ihlenfeldt, "Benchmarking automl for regression tasks on small tabular data in materials design," *Scientific Reports*, vol. 12, no. 1, p. 19350, 2022.

[17] S. B. Rabbani, I. V. Medri, and M. D. Samad, "Attention versus contrastive learning of tabular data - A data-centric benchmarking," *CoRR*, vol. abs/2401.04266, 2024.

[18] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep neural networks and tabular data: A survey," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 35, no. 6, pp. 7499–7519, 2024.

[19] Z. Qian, R. Davis, and M. van der Schaar, "Synthcity: a benchmark framework for diverse use cases of tabular synthetic data," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[20] N. Erickson, L. Purucker, A. Tschalzev, D. Holzmüller, P. M. Desai, F. Hutter *et al.*, "Tabarena: A living benchmark for machine learning on tabular data," *arXiv preprint arXiv:2506.16791*, 2025.

[21] M. A. Oliver and R. Webster, "Basic steps in geostatistics: the variogram and kriging," Springer, Tech. Rep., 2015.

[22] M. Oliver and R. Webster, "A tutorial guide to geostatistics: Computing and modelling variograms and kriging," *Catena*, vol. 113, pp. 56–69, 2014.

[23] C. Brunsdon, S. Fotheringham, and M. Charlton, "Geographically weighted regression," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 47, no. 3, pp. 431–443, 1998.

[24] A. S. Fotheringham, W. Yang, and W. Kang, "Multiscale geographically weighted regression (mgwr)," *Annals of the American Association of Geographers*, vol. 107, no. 6, pp. 1247–1265, 2017.

[25] Q. Zhu and H. Lin, "Comparing ordinary kriging and regression kriging for soil properties in contrasting landscapes," *Pedosphere*, vol. 20, no. 5, pp. 594–606, 2010.

[26] V. Van Zoest, F. B. Osei, G. Hoek, and A. Stein, "Spatio-temporal regression kriging for modelling urban no2 concentrations," *International Journal of Geographical Information Science*, vol. 34, no. 5, pp. 851–865, 2020.

[27] S. Araki, K. Yamamoto, and A. Kondo, "Application of regression kriging to air pollutant concentrations in japan with high spatial resolution," *Aerosol and Air Quality Research*, vol. 15, no. 1, pp. 234–241, 2015.

[28] M. P. Lucas, R. J. Longman, T. W. Giambelluca, A. G. Frazier, J. Mclean, S. B. Cleveland, Y.-F. Huang, and J. Lee, "Optimizing automated kriging to improve spatial interpolation of monthly rainfall over complex terrain," *Journal of Hydrometeorology*, vol. 23, no. 4, pp. 561–572, 2022.

[29] B. Huang, B. Wu, and M. Barry, "Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices," *International Journal of Geographical Information Science*, vol. 24, no. 3, pp. 383–401, 2010.

[30] Z. Zhu, B. Li, Y. Zhao, Z. Zhao, and L. Chen, "Socio-economic impact mechanism of ecosystem services value, a pca-gwr approach," *Polish Journal of Environmental Studies*, vol. 30, no. 1, pp. 977–986, 2020.

[31] S. Li, Z. Zhao, X. Miaomiao, and Y. Wang, "Investigating spatial non-stationary and scale-dependent relationships between urban surface temperature and environmental factors using geographically weighted regression," *Environmental Modelling & Software*, vol. 25, no. 12, pp. 1789–1800, 2010.

[32] P. Harris, A. Fotheringham, R. Crespo, and M. Charlton, "The use of geographically weighted regression for spatial prediction: an evaluation of models using simulated data sets," *Mathematical Geosciences*, vol. 42, pp. 657–680, 2010.

[33] S. Kumar, R. Lal, and D. Liu, "A geographically weighted regression kriging approach for mapping soil organic carbon stock," *Geoderma*, vol. 189, pp. 627–634, 2012.

[34] M. Imran, A. Stein, and R. Zurita-Milla, "Using geographically weighted regression kriging for crop yield mapping in west africa," *International Journal of Geographical Information Science*, vol. 29, no. 2, pp. 234–257, 2015.

[35] B. S. Murphy, "Pykrige: development of a kriging toolkit for python," in *AGU Fall Meeting Abstracts*, vol. 2014, 2014, pp. H51K–0753.

[36] S. Müller, N. Hollmann, S. P. Arango, J. Grabocka, and F. Hutter, "Transformers can do bayesian inference," *arXiv preprint arXiv:2112.10510*, 2021.

[37] R. M. Neal, *Bayesian learning for neural networks*. Springer Science & Business Media, 2012, vol. 118.

[38] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, ser. JMLR Workshop and Conference Proceedings, vol. 48. JMLR.org, 2016, pp. 1050–1059.

[39] J. Pearl, *Causality*. Cambridge University Press, 2009.

[40] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[41] B. Peters, V. Niculae, and A. F. T. Martins, "Sparse sequence-to-sequence models," in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Association for Computational Linguistics, 2019, pp. 1504–1519.

[42] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021, pp. 18 932–18 943.

[43] J. Chen, K. Liao, Y. Wan, D. Z. Chen, and J. Wu, "Danets: Deep abstract networks for tabular data classification and regression," in *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 2022, pp. 3930–3938.

[44] L. Anselin, R. Bongiovanni, and J. Lowenberg-DeBoer, "A spatial econometric approach to the economics of site-specific nitrogen management in corn production," *American Journal of Agricultural Economics*, vol. 86, no. 3, pp. 675–687, 2004.

[45] I. O. Odeh, A. McBratney, and D. Chittleborough, "Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging," *Geoderma*, vol. 67, no. 3-4, pp. 215–226, 1995.

[46] T. Hengl, G. B. Heuvelink, and A. Stein, "A generic framework for spatial prediction of soil variables based on regression-kriging," *Geoderma*, vol. 120, no. 1-2, pp. 75–93, 2004.

[47] T. Hengl, G. B. Heuvelink, and D. G. Rossiter, "About regression-kriging: From equations to case studies," *Computers & geosciences*, vol. 33, no. 10, pp. 1301–1315, 2007.

[48] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[49] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, ser. COLT '92. New York, NY, USA: Association for Computing Machinery, Jul. 1992, pp. 144–152.

[50] M. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, Jul. 1998.

[51] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge: Cambridge University Press, 2000.

[52] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[53] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[54] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[55] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[56] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Inf. Fusion*, vol. 81, pp. 84–90, 2022.

[57] M. Joseph and H. Raj, "Gandalf: gated adaptive network for deep automated learning of features," *arXiv preprint arXiv:2207.08548*, 2022.

[58] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.
[59] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017, pp. 4765–4774.