

**International Journal on
Advances in Systems and Measurements**



The *International Journal on Advances in Systems and Measurements* is published by IARIA.

ISSN: 1942-261x

journals site: <http://www.ariajournals.org>

contact: petre@aria.org

Responsibility for the contents rests upon the authors and not upon IARIA, nor on IARIA volunteers, staff, or contractors.

IARIA is the owner of the publication and of editorial aspects. IARIA reserves the right to update the content for quality improvements.

Abstracting is permitted with credit to the source. Libraries are permitted to photocopy or print, providing the reference is mentioned and that the resulting material is made available at no cost.

Reference should mention:

International Journal on Advances in Systems and Measurements, issn 1942-261x
vol. 17, no. 3 & 4, year 2024, http://www.ariajournals.org/systems_and_measurements/

The copyright for each included paper belongs to the authors. Republishing of same material, by authors or persons or organizations, is not allowed. Reprint rights can be granted by IARIA or by the authors, and must include proper reference.

Reference to an article in the journal is as follows:

<Author list>, "<Article title>"
International Journal on Advances in Systems and Measurements, issn 1942-261x
vol. 17, no. 3 & 4, year 2024, http://www.ariajournals.org/systems_and_measurements/

IARIA journals are made available for free, proving the appropriate references are made when their content is used.

Sponsored by IARIA

www.aria.org

Copyright © 2024 IARIA

Editors-in-Chief

Constantin Paleologu, University "Politehnica" of Bucharest, Romania
Sergey Y. Yurish, IFSA, Spain

Editorial Board

Nebojsa Bacanin, Singidunum University, Serbia
Chaity Banerjee, University of Alabama in Huntsville, USA
Robert Bestak, Czech Technical University in Prague, Czech Republic
Michał Borecki, Warsaw University of Technology, Poland
Vitor Carvalho, 2Ai | School of Technology | IPCA & Algoritmi Research Center | Minho University, Portugal
Paulo E. Cruvinel, Brazilian Corporation for Agricultural Research (Embrapa), Brazil
Miguel Franklin, Federal University of Ceara, Brazil
Mounir Gaidi, University of Sharjah, UAE
Eva Gescheidtova, Brno university of Brno, Czech Republic
Franca Giannini, CNR - Istituto di Matematica Applicata e Tecnologie Informatiche "Enrico Magenes", Italy
Terje Jensen, Telenor, Norway
Wooseong Kim, Gachon University, South Korea
Dragana Krstic, University of Nis, Serbia
Andrew Kusiak, The University of Iowa, USA
Diego Liberati, CNR-IEIT, Italy
D. Manivannan, University of Kentucky, USA
Stefano Mariani, Politecnico di Milano, Italy
Constantin Paleologu, National University of Science and Technology Politehnica Bucharest, Romania
Paulo Pinto, Universidade Nova de Lisboa, Portugal
R. N. Ponnalagu, BITS Pilani Hyderabad campus, India
Leon Reznik, Rochester Institute of Technology, USA
Gerasimos Rigatos, Unit of Industrial Automation - Industrial Systems Institute, Greece
Claus-Peter Rückemann, Universität Münster / DIMF / Leibniz Universität Hannover, Germany
Subhash Saini, NASA, USA
Adérito Seixas, Escola Superior de Saúde Fernando Pessoa, Porto, Portugal
V. R. Singh, National Physical Laboratory (NPL), New Delhi, India
Miroslav Velez, Aries Design Automation, USA
Manuela Vieira, Instituto Superior de Engenharia de Lisboa (ISEL), Portugal
Xianzhi Wang, University of Technology Sydney, Australia
Kaidi Wu, College of Mechanical Engineering | Yangzhou University, China
Linda Yang, University of Portsmouth, UK
Sergey Y. Yurish, IFSA, Spain
Daniele Zonta, University of Trento / National Research Council, Italy

CONTENTS

pages: 83 - 95

Meet DebiAI: A Versatile Open-Source Tool for Streamlined Data Analysis, Visualization, and ML Model Evaluation

Tom Mansion, IRT SystemX, France
Raphaël Braud, IRT SystemX, France
Faouzi Adjed, IRT SystemX, France
Ahmed Amrani, IRT SystemX, France
Sabrina Chaouche, IRT SystemX, France
Fady Bekkar, IRT SystemX, France
Yoann Randon, IRT SystemX, France
Martin Gonzalez, IRT SystemX, France
Loïc Cantat, IRT SystemX, France

pages: 96 - 110

Development of Children's Crossing Skills in Urban Area: Visual Exploration and Mental Representation about Hazards

Jordan Solt, Lorraine University, 2LPN, Luxembourg
Jerome Dinet, Lorraine University and Chair BEHAVIOUR, France
Muneo Kitajima, Nagaoka University of Technology, Japan
Aurelie Mailloux, Reims hospital, URCA, France
Samuel Ferreira Da Silva, Lorraine University and Chair BEHAVIOUR, France
Gaelle Nicolas, Lorraine University and Chair BEHAVIOUR, France

pages: 111 - 126

Caption Generation for Clothing Image Pair Comparison Using Attribute Prediction and Prompt-based Visual Language Model

Soichiro Yokoyama, Hokkaido University, Japan
Kohei Abe, Hokkaido University, Japan
Tomohisa Yamashita, Hokkaido University, Japan
Hidenori Kawamura, Hokkaido University, Japan

pages: 127 - 137

Day-ahead Forecasting Electricity Spot Prices in Norway with ARIMA, XGBoost, and LSTM Models

Markus Jensen, Kristiania University College, Norway
Huamin Ren, Kristiania University College, Norway
Andrii Shalaginov, Kristiania University College, Norway

pages: 138 - 145

Improving Effectiveness and Performance Based on Dimensionality Reduction of CCD Image Features in Fall Armyworm's Control

Alex Bertolla, Embrapa Instrumentation, and Federal University of São Carlos - Post Graduation Program in Computer Science, Brazil
Paulo Cruvinel, Embrapa Instrumentation, and Federal University of São Carlos - Post Graduation Program in Computer Science, Brazil

pages: 146 - 155

Evaluating Digital Avatars - A Systematic Approach to Quantify the Uncanny Valley Effect by Using Real Life Samples

Hakan Arda, Faculty of Computer Science and Business Information Systems Technical University of Applied Sciences Würzburg-Schweinfurt, Germany

Andreas Henneberger, Faculty of Computer Science and Business Information Systems Technical University of Applied Sciences Würzburg-Schweinfurt, Germany

Karsten Huffstadt, Faculty of Computer Science and Business Information Systems Technical University of Applied Sciences Würzburg-Schweinfurt, Germany

Nicholas Müller, Faculty of Computer Science and Business Information Systems Technical University of Applied Sciences Würzburg-Schweinfurt, Germany

pages: 156 - 165

AI-based Automated Production of Learning Content – A Means to Bridging the Digital Divide in Workplace Learning?

Katharina Frosch, Brandenburg University of Applied Sciences, Deutschland

Friederike Lindauer, Brandenburg University of Applied Sciences, Germany

pages: 166 - 175

Developing a Sign Language Writing System: Focus on Necessity and Sign Language-Specific Features

Nobuko Kato, Tsukuba University of Technology, Japan

Yuito Nameta, Tsukuba University of Technology, Japan

Megumi Shimomori, Tsukuba University of Technology, Japan

Akihisa Shitara, University of Tsukuba, Japan

Sumihiro Kawano, Tsukuba University of Technology, Japan

Yuhki Shiraishi, Tsukuba University of Technology, Japan

pages: 176 - 188

Regression Model-Based Prediction for Building Energy Star Score of New York City

Fan Zhang, Tuskegee University, United States

Baiyun Chen, Tuskegee University, United States

Faria Brishti, Tuskegee University, United States

Sameeruddin Mohammed, Tuskegee University, United States

Fan Wu, Tuskegee University, United States

Ling Bai, The University of British Columbia, Canada

pages: 189 - 200

Symbolic Unfolding versus Tuning of Similarity-based Fuzzy Logic Programs

Gines Moreno, University of Castilla-La Mancha, Spain

José Antonio Riaza, University of Castilla-La Mancha, Spain

pages: 201 - 208

Evaluation of AI Learning Materials Using Physical Computing

Toshiyasu Kato, Nippon Institute of Technology, Japan

Yuto Chino, Fuchu City Fuchu 6th Junior High School, Japan

Meet DebiAI: A Versatile Open-Source Tool for Streamlined Data Analysis, Visualization, and ML Model Evaluation

Tom Mansion

IRT SystemX

91120 Palaiseau, France
tom.mansion@irt-systemx.fr

Raphaël Braud

IRT SystemX

91120 Palaiseau, France
raphael.braud@irt-systemx.fr

Faouzi Adjed

IRT SystemX

91120 Palaiseau, France
faouzi.adjed@irt-systemx.fr

Ahmed Amrani

IRT SystemX

91120 Palaiseau, France
ahmed.amrani@irt-systemx.fr

Sabrina Chaouche

IRT SystemX

91120 Palaiseau, France
sabrina.chaouche@irt-systemx.fr

Fady Bekkar

IRT SystemX

91120 Palaiseau, France
fady.bekkar@irt-systemx.fr

Yoann Randon

IRT SystemX

91120 Palaiseau, France
yoann.randon@irt-systemx.fr

Martin Gonzalez

IRT SystemX

91120 Palaiseau, France
martin.gonzalez@irt-systemx.fr

Loïc Cantat

IRT SystemX

91120 Palaiseau, France
loic.cantat@irt-systemx.fr

Abstract—We present DebiAI, a powerful open source tool crafted to streamline data analysis, visualization, and the comprehensive evaluation and comparison of Machine Learning models. It serves as a versatile companion throughout the entire machine learning workflow, from project data preparation to model performance assessment. With its intuitive and feature-rich graphical interface, DebiAI enables users to effortlessly visualize, explore, select, edit, and annotate both data and metadata. The tool is also equipped for bias detection and contextual evaluation of ML models, ensuring a thorough and fair analysis. Built on a flexible, generic data model, DebiAI is adaptable to a wide range of ML tasks, including classification, regression, and object detection in images, as well as a variety of tasks for time-series and more. Released under the Apache License, Version 2.0, it offers an accessible and linearly scalable solution for ML practitioners of all levels. The code for the proposed tool is publicly available at <https://github.com/debiai>; and other information and user guidelines are available on the dedicated website: <https://debiai.irt-systemx.fr>.

Index Terms—Data Analysis; Data Visualization; Bias Detection; Human-Centered Machine Learning; Trustworthy AI.

I. INTRODUCTION

This work is an extension of our previous work published in the ICAS conference [1] dealing with data analysis and visualization in Machine Learning (ML) projects. They are playing a crucial role in a typical ML process, and they are not only contributing in the data preparation phase, but also during and after the model building. In short, data visualization contributes to the whole ML life cycle, from its specifications and data acquisitions until its deployment and monitoring. This involved to create an emerging research topic, which

combines several interactive systems and domains for ML processes, focused on human interaction and collaboration [2] to constitute a new field that is named Human-Centered Machine Learning (HCML) or Human-Centered Artificial Intelligence (HCAI) interaction [3], [4]. Thus, a typical HCML framework allows an interactive visual analysis and evaluation of data and ML models [5]. As a result, efficient tools are essential to support users in the most user-friendly manner throughout the entire ML process, including tasks like data preparation and quality inspection prior to training, as well as monitoring model performance and deployment quality after training. Specifically, as highlighted by Caple et al. [4], the HCML/HCAI process encompasses five key areas: (i) Explainable and Interpretable Artificial Intelligence (AI), (ii) Human-Centered Approaches to AI Design and Evaluation, (iii) Human-AI Collaboration, (iv) Ethical AI, and (v) AI Interaction.

An effective tool should support the iterative ML process across multiple stages: from data preparation, analysis, anomaly detection, and annotation, to the evaluation and analysis of model results and performance. This helps identify model weaknesses and uncover issues at the data level. In real-world ML projects, such as those in industry, data is often enriched with metadata, including operational context and expert knowledge, which provides deeper insights into the raw data and enhances the learning process. This, in turn, improves the quality of model training and predictions. Furthermore, having such tools increases the trustworthiness of the ML algorithms in use.

In ML-based engineering systems, it is crucial to guarantee key properties like accuracy, robustness, explainability,

Research funded by the French Government under the “France 2030” program.

fairness, privacy, among many other primary values of AI trustworthiness. Current research and development challenges of deploying trustworthy ML solutions are covered by wide programs such as Confiance.ai [6], the French AI flagship program to industrialize trustworthy AI-based critical systems [7], [8] and the TAILOR [9] network at the European level.

DebiAI has been developed by the IRT SystemX in the framework of Confiance.ai program to contribute in ensuring trustworthiness by data, and serves as the main interface to view, analyze, select, edit and/or annotate any type of data and metadata.

The remainder of the paper is structured as follows. Section II provides a brief review of the state-of-the-art HCML tools. Section III outlines the methodology developed in this work, building on our previous contribution [9]. Section IV details the implementation, presenting the main architecture and various functionalities. In Section V, we describe the application of the tool to real-world use cases, including images and time series. Section VI discusses evaluation analysis and usage recommendations, while Section VII addresses limitations. Finally, conclusions and future perspectives are presented in Section VIII.

II. LITERATURE REVIEW

Data visualization is the practice of representing information using graphical representations, employing technology-driven tools and software. Its fundamental objective is to enhance pattern recognition, improve understanding of complex concepts and facilitate in-depth exploration, thereby fostering the generation of new insights. Well-designed data visualizations can help in understanding large datasets and establishing connections between ideas, concepts, and processing stages. Therefore, visual analysis can contribute to the optimization of AI approaches by actively participating in all aspects of the model building process [10], [11]. Similarly, Hohman et al. [12] highlight that successful ML applications often require iterations in data handling and continuous adjustments of the model. The authors introduced CHAMELEON, an interactive tool designed to attribute data iteration, thereby enhancing model performance, data validation, and the overall quality of ML projects. To facilitate the interaction between machine learning experts and final users, Françoise et al. [13] proposed a toolkit addressing an interactive machine learning workflow to permit a collaboration between machine learning experts, designers and end-users through a unified tool. To improve data quality, Kandel et al. [14] presented Profiler, a tool using data mining to automatically detect issues and recommend coordinated visualizations for context-based assessment. Profiler offers methods for integrated statistical and visual analysis and view suggestions.

Grafana [15] is an open source web platform, in its recent version, used for data visualization and tracking in real time for data science field. This tool has multiple functionalities for data monitoring, whether user's data is connected to Prometheus [16], InfluxDB [17] and others database. Thus, the platform offers an increasing array of data analysis and

generative AI features, such as creating alerts, predicting capacity needs, and detecting anomalous activities [18]–[20]. Streamlit [21] is a Python library that enables data scientists to swiftly and effectively create robust web applications. The advantage of Streamlit lies in its simplicity of using python scripts to build customizable dashboards, whereas DebiAI has the advantage of offering a wide range of ready to use widgets available.

ScrutinAI [22] is a Visual Analytics tool specifically tailored for enhancing the comprehension of deep neural network (DNN) predictions. Its primary objective is to identify and investigate potential weaknesses within models. To facilitate this, ScrutinAI provides interactive visualizations of input and output data, along with interactive plots and data filtering for comprehensive analysis of predictions. This tool is specifically designed for object detection and semantic segmentation, whereas DebiAI is applicable to a wide range of use cases. Zhang et al. [23] presented Manifold, a visual analytics platform designed for comparing and debugging ML models. The platform empowers users to categorize instances based on the model's accuracy and confidence, identify symptomatic instances that generate incorrect results and continually help to enhance the model's performance. As DebiAI, Manifold is created as a generic tool that operates independently of the internal logic of the ML model. It focuses on the input and the output. Similarly, Uni-Evaluator [24] is also independent of the model but focuses on evaluating computer vision tasks, such as classification, object detection, and instance segmentation, with tailored visualizations for these specific uses cases. However, DebiAI can perform a more general evaluation across different types of ML tasks thanks to its flexible data model.

To improve model performance and understand their limitations, it is not enough to just rely on the overall results from the test and training sets. To overcome this limitation, the ModelTracker [25] tool provides instance-level result visualization, allowing users to inspect each instance individually. The tool has been applied to the binary classification task, and in [26], the authors introduced Squares, which extends the approach to multi-class classification. Additionally, Squares facilitates the estimation of common performance metrics and provides instance-level result visualization, guiding practitioners in troubleshooting performance issues while offering direct access to relevant data. They used Parallel Coordinates Plots (PCP) to visually represent the multi-class predictions for a subset of instances. In line with this approach and to improve the understanding of models results, the proposed DebiAI extends the analysis by enabling the exploration of model outcomes at various levels of granularity, including instance, subset, and dataset. This functionality has been applied to multiple tasks such as regression, classification, object detection, etc. In the same way as outlined in [26], DebiAI utilizes PCP to analyze model results. The implementation of PCP within DebiAI is flexible, enabling its use not only for result analysis but also for assessing attributes. Table I summarizes an overview of HCML tools presented in this section by

TABLE I: Summary of Tools in AI and Data Visualization

Tool	Purpose	Key Features
CHAMELEON [12]	Enhances ML model performance and data quality through iteration	Interactive tool for data attribution, model performance enhancement, and data validation
Marcelle [13]	Facilitates interaction between ML experts, designers, and end-users	Unified toolkit for interactive machine learning workflows
Profiler [14]	Improves data quality using data mining and visualizations	Integrated statistical and visual analysis, view suggestions for context-based assessment
Grafana [15]	Real-time data visualization and tracking for data science	Supports multiple databases (e.g., Prometheus, InfluxDB), alerting, predictive analysis, and anomaly detection
Streamlit [21]	Simplifies creation of data-driven web applications	Python-based, customizable dashboards, fast prototyping
ScrutinAI [22]	Enhances understanding of DNN predictions	Interactive visualizations, data filtering, tailored for object detection and semantic segmentation
Manifold [23]	Compares and debugs ML models	Instance categorization based on accuracy/confidence, generic tool for input-output analysis
Uni-Evaluator [24]	Evaluates computer vision tasks	Tailored visualizations for classification, object detection, and instance segmentation
ModelTracker [25]	Provides instance-level result visualization for binary classification	Instance-level inspection and analysis
Squares [26]	Extends ModelTracker to multi-class classification	Instance-level result visualization using PCP
DebiAI	General evaluation across different ML tasks	Interactive instance, subset, and dataset-level analysis, flexible PCP implementation for results and attributes analysis

highlighting their purposes and their main key features.

III. METHODOLOGY

DebiAI is a web-based visual analytics application designed to support ML and data analysis. Its emphasis lies in two crucial phases of the ML pipeline: pre-model and post-model building. As shown in Fig. 1, DebiAI facilitates the development of ML models by assisting in data analysis during data curating and processing stage and enabling models performances comparison.

In the pre-model construction phase, DebiAI serves as a key resource for data scientists and ML engineers during project preparation. It enables them to visually identify biases and errors in data inputs, detect anomalies and outliers throughout the data life cycle, assess data quality and domain coverage through relevant metrics and select and analyze subsets of data to improve the quality of ML models.

In the post-model building phase, DebiAI serves as a visual analytics solution, simplifying the interpretation of the ML

model's outputs. Its primary objective is to present the model's results in an intuitive and easily understandable manner, ultimately enhancing user confidence in the model's predictions. Additionally, DebiAI offers features to identify model's weaknesses, comparing performances, and evaluating model's effectiveness according to the project contextual data such as weather or gender biases. This comprehensive approach fosters ongoing model refinement, tailored to the specific needs of the use case.

In both phases, DebiAI provides tools for creating and sharing statistical visualizations of the project data and results with collaborators (team or/and clients). In summary, DebiAI reflects its name from "Debiasing AI" towards mitigate AI bias [27].

A. Functional Description

DebiAI is an intuitive visualization tool designed to simplify the creation of interactive dashboards, empowering users without little to no programming skills. It offers a diverse set of

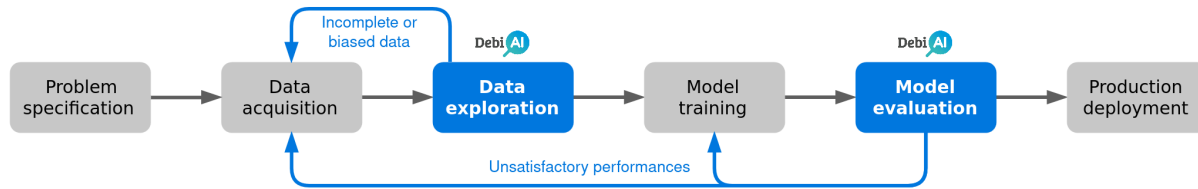


Fig. 1: The role of DebiAI within the different stages of an AI project. This diagram illustrates a comprehensive machine learning pipeline from problem specification to production deployment. DebiAI intervenes during the data exploration stage and the contextual comparison stage of AI project models. It helps provide intelligence on incomplete or biased data and on model performance, thereby accelerating the resolution of feedback loops.

graphical widgets, including charts, tables, parallel categories, parallel coordinate plots, interval plots, night star plots, and sample arrays. Moreover, DebiAI provides a user-friendly and flexible solution for interactive dashboard design, allowing users to effortlessly configure, adjust, resize, and position these widgets within their dashboards, ensuring the utmost customization of data presentation. This includes the ability to generate and share statistical visualizations of project data with team members or clients, fostering collaboration and informed decision making by providing clear insights into the data.

One of DebiAI's standout features is its dynamic data selection and filtering capabilities, which encourages continuous exploration. Users can effortlessly create data subsets (selections) and apply filters based on various variables and contexts. This ensures that the dashboard consistently presents the identified subset of data. Furthermore, DebiAI assists users in identifying biases and inaccuracies in inputs, results, project data contexts, or ground truths, thereby improving data integrity.

DebiAI facilitates the evaluation and comparison of ML model's performances within the whole dataset or a specific data subset. It enables analysis of results across multiple levels of granularity. Indeed, the model's performances are calculated at the level of each instance. Consequently, it is possible to identify the contexts or a combination of contexts in which the model encounters difficulties. It also simplifies the generation and organization of datasets, supporting in-depth analysis and potential retraining.

DebiAI relies on a generic data model that facilitates seamless application across various datasets, data types, and use cases while maintaining consistent data processing practices. This essential feature provides DebiAI with flexibility, allowing it to transition between various datasets or the results of the model. In addition to its visualization capabilities, DebiAI incorporates implementations of statistical measures such as correlation analysis using Pearson or Spearman coefficients. To support these visualizations, DebiAI also integrates techniques for discretizing continuous variables. In addition, it enables the use of internal or external algorithms to compute metrics or indicators on the data. Consequently, these metrics can be calculated either before integrating the data into DebiAI or during the data analysis phase. Various types of calculations can potentially be carried out by these algorithms, including

the computation of new features, the assessment of model's results quality, as well as indicators of data quality and distribution. In addition to that, at each step, DebiAI provides an easy solution to transform the dashboard of widgets and comments into a markdown file with a PNG image format for each widget.

IV. IMPLEMENTATION DESCRIPTION

In this section, we describe DebiAI's global architecture and dive into the details of each component. We start by an overview of the architecture details, then we proceed by presenting the data model, generic and multi-dimensional, followed by data integration process.

A. DebiAI Technical architecture

The DebiAI architecture is divided into two main environments as shown in Fig. 2, which are the project environment (data and algorithms) and the application environment (back-end and dashboard).

The project environment consists on the following:

- **Project Data:** This is the source of data that the user intends to analyze. It may originate from various sources and formats, such as CSV or JSON.
- **Data-Providers:** These are the services created by the project members to enable DebiAI to fetch data and model results directly from the project's sources. Creating a Data-provider allows DebiAI to always fetch up-to-date data without duplication. A Web Data-Provider can be developed using any programming language, access data from any type of database, and be hosted on any server. The only stipulation is that it should implement and expose a specific REST API according to a defined contract. DebiAI allows users to add as many Data-Providers as they require, allowing them to analyze different projects with mixed data sources.
- **Python Scripts and DebiAI Python Module:** Using the DebiAI Python module, users can adapt their existing scripts and workflows to create selections and insert data and model's results into DebiAI. This Python module is a simpler alternative to creating a Data-provider, but it requires data to be duplicated and the module to be called at each project data update, which is time consuming.

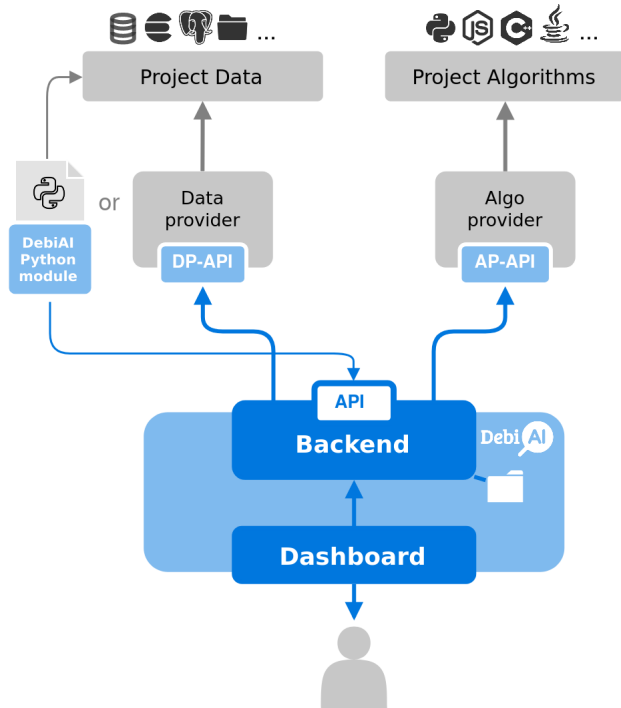


Fig. 2: DebiAI system architecture overview illustrating the division between the project environment, where users manage their data sources, data providers, algorithms, algorithm providers, and the DebiAI application environment, which includes the backend, web dashboard, and data storage. The DebiAI Python module serves as a simple way to insert data into DebiAI. This architecture enables seamless integration of multiple and mixed data sources, real-time data updates for project analysis, and project algorithms.

The comparison with Data-provider services is done in Section VI.

- Algo-Providers:** These services are used to provide specific algorithms required by projects. Once an algorithm is provided to DebiAI, it can be called from the analysis dashboard with the project's data. For example, an algorithm can be used to compute some specific features, model prediction, data quality metrics, etc. The algorithm's results can be displayed, filtered, and analyzed, just like any other dataset. An Algo-Provider can be developed using any programming language, expose any algorithm and be hosted on any server. The only stipulation is that it should implement and expose a specific REST API according to a defined contract. Users can add as many Algo-Providers as they require.

The DebiAI's application environment consist on the following:

- Backend and API:** This is a Python-powered backend that provides an API and serves the Web dashboard. This API is employed by the dashboard for data retrieval and by the Python module for data insertion. Additionally, it manages communications with the Web Data-providers,

processes computational requests made by the dashboard, and calls the Algo-Providers selected from the dashboard.

- DebiAI Web Dashboard:** This is the user interface of DebiAI, developed using VueJs. It provides users with an interactive platform to manage and view their data, and is hosted and served by the DebiAI backend. DebiAI uses different tools to display plots, the main being the PlotlyJs library.
- Data storage:** DebiAI uses a folder-based data store that contains data in a JSON format. This data store supports the DebiAI backend by retaining projects created by the Python module and some specific dashboard elements, including layout configurations for project dashboards.

B. DebiAI Generic Data Model

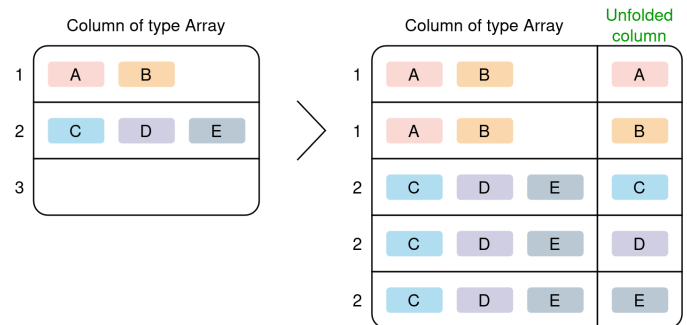


Fig. 3: Vertical Unfolding : DebiAI allows array columns to be **unfolded vertically**, on the condition that the data follows certain formatting conditions. This figure exposes how unfolding an array column vertically will add more lines. Note that because the array for data number 3 is empty, the value for data number 3 is absent from the unfolded column.

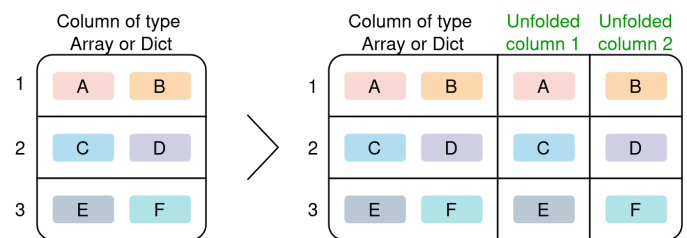


Fig. 4: Horizontal Unfolding: DebiAI allows array and dictionaries columns to be **unfolded horizontally**, on the condition that the data follows certain formatting conditions. This figure exposes how unfolding a dictionary or array column horizontally will add new columns.

One of the most important features of DebiAI is its data model. The main objective is to enable the determination of the format of instances and the relationship between instances, models, models' outputs, and models' evaluation metrics per instance. Syntactically, each instance is composed of attributes, contexts, and annotations. The instance is linked to multiple ML models, where each model produces an output. Evaluation metrics are also associated with the model's outputs.

- 1) **Data purpose:** This structure is applicable to all types of data and ML tasks such as classification, regression, object detection, anomaly detection, previsions.
- 2) **Data format:** The data format required must follow a CSV format and/or table structures. Dataframe and NumPy arrays are supported by the DebiAI data insertion Python module. When supplying data through a data-provider (described in Fig. 2), any format is acceptable since it follows the requirements of the data-provider's API.
- 3) **Primitive data type:** DebiAI supports primitive types such as text, numbers and boolean values. Any data columns with text will be considered as a class column. Class columns are treated differently in DebiAI, for example, the distribution plot will set the bins number to the number of unique values in the class columns, but set a fixed bin number for number columns. Columns with only numerical data can be forced into a class type, this can be useful when numeric values needs to be considered as a class, such as vehicle model number, a year date or age.
- 4) **Missing data:** Missing, None or NaN data are supported by DebiAI since version 0.29.0. The percentage of missing values is displayed for each column, and missing values can be filtered out or in.
- 5) **Arrays and lists data:** DebiAI supports columns of values containing lists under the following condition: the column must contain only lists. If the condition is met, DebiAI will be able to unfold the list of values as new lines, changing the scope of the analysis (as explained in Fig. 3). This process is called "vertical unfolding", it is demonstrated in Section IV-C. If all the columns list values have the same number of keys, the horizontal unfolding is available. Unfolding the lists horizontally will treat the list values as individual columns (as explained in Fig. 4). If the list recursively contains more lists, the unfolding operation can be repeated. However, columns containing some dictionaries and some other mixed types won't be able to be unfolded.
- 6) **Dictionaries and JSON objects data:** DebiAI supports columns values containing dictionaries (a data element composed of values associated with keys) under certain conditions: the column must contain only dictionaries, the dictionaries must have the same keys for all values, the number of keys must not exceed 30. If these conditions are met, DebiAI will be able to unfold the dictionaries values and treat them as individual columns (as explained in Fig. 4). If the dictionaries recursively contain more dictionaries, the unfolding operation can be repeated. However, as vertical unfolding, columns containing other types than dictionaries or dictionaries with different keys won't be able to be unfolded.

C. Multi-Dimensional Data Model in DebiAI

DebiAI's data model is equipped to handle complex, multidimensional datasets, making it particularly valuable in

projects like the Woodscape dataset [28] where both images and the objects are connected (as illustrated in Fig. 5): within them are analyzed. The model supports recursive and nested objects, such as lists and dictionaries, allowing users to progressively explore and **unfold** various data dimensions. In projects like Woodscape, the dataset initially comprises 1,624 images. By unfolding these images into their annotated objects, the analysis scope broadens, covering 72,061 objects.



Fig. 5: Hierarchical structure of the Woodscape dataset: Layer1 represents images, while Layer2 represents annotated objects within those images.

This feature enables a detailed analysis of specific aspects, such as the distribution of object classes (e.g., vehicles, pedestrians) and position of the objects across the dataset. Such a capability is essential for projects with hierarchically structured data, facilitating a comprehensive analysis at multiple levels. In this multidimensional model, users can start with a high-level analysis of the entire dataset, then delve deeper into specific annotations, like object detection labels. This process provides valuable insights into the distribution and characteristics of different object classes, helping to refine model training and evaluation strategies.

Furthermore, DebiAI enhances its selection capabilities by ensuring that any selection made on the unfolded data (such as the Woodscape objects) will also select the corresponding original data (such as the Woodscape images as illustrated in Fig. 7). This feature greatly improves the accuracy and relevance of data selection, allowing users to maintain a consistent context when analyzing both high-level and detailed data aspects. This multi-layered approach enhances the depth and breadth of data analysis, allowing for the identification of biases or gaps in the dataset. DebiAI's capability to analyze data at various levels of granularity makes it an indispensable tool for developing trustworthy AI systems, ensuring a thorough and nuanced understanding of ML datasets.

D. DebiAI Data Integration Process

DebiAI offers two main ways to add data, each suited for different types of users and projects:

- 1) **Python Module:** This principal method enables seamless integration of project's data into DebiAI via a dedicated Python module. Made for an integration within Python workflows, this approach, for example, facilitates the direct transfer of models' results post-evaluation. This method is especially handy for those who primarily use Python.

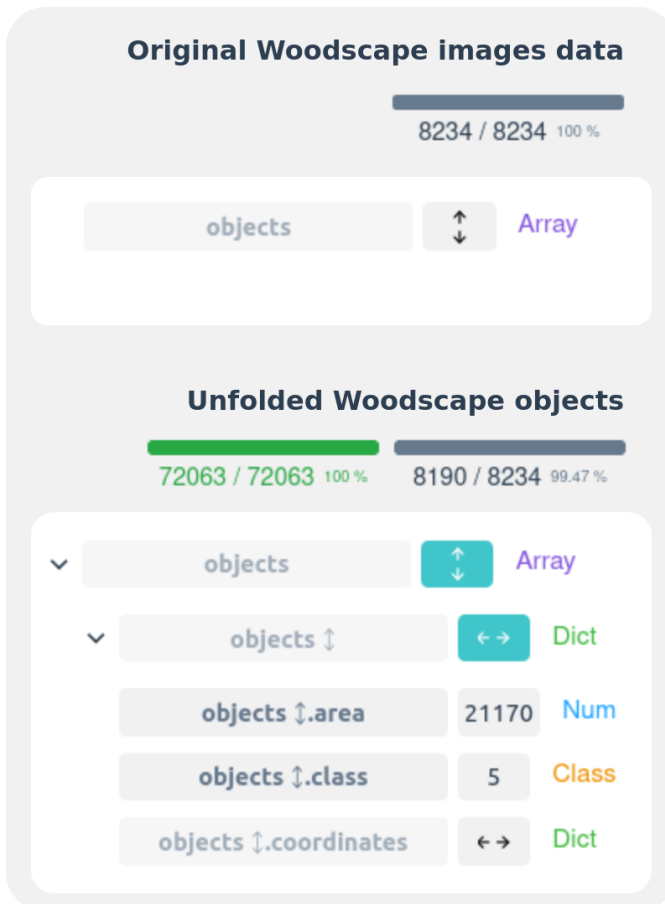


Fig. 6: DebiAI interface showing the number of images and objects in the Woodscape dataset, before and after unfolding the object layer. New columns are available, increasing the analysis depth. Note that 44 (0.53%) images aren't selected after unfolding due to the absence of objects in those images.

- 2) **Data Providers:** Alternatively, DebiAI can interface with data through RESTful services, termed 'Data Providers'. This method is database-agnostic, allowing DebiAI to directly request project's data, thereby making the data loading process faster and more efficient. Unlike the Python module, it doesn't require DebiAI to duplicate data within its integrated database. Although setting up a Data Provider is more time-consuming than using the Python module, it offers greater efficiency and flexibility. This is particularly beneficial for long-term projects that regularly update their data.

Each method offers distinct benefits, and the choice depends on the specific requirements and scale of the project.

V. DEBIAI APPLICATION

DebiAI is built upon a generic data model, and does not depend on data type (for instance images and time series), making it pertinent to various use cases across a multitude of datasets. This intrinsic adaptability allows it to be valuable in a wide range of scenarios. It demonstrates its utility in the

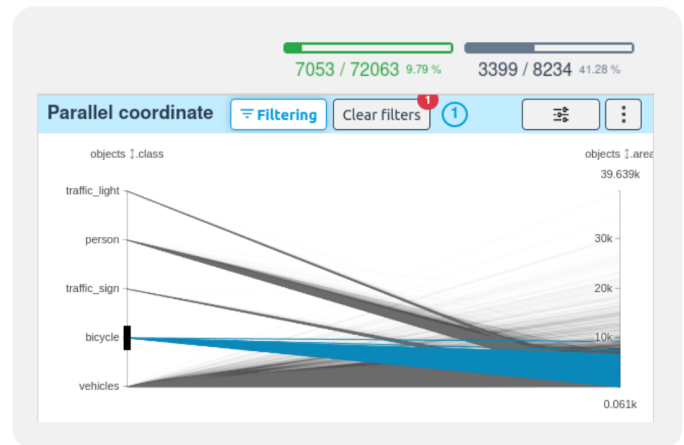


Fig. 7: DebiAI interface showing the selection of objects of type "bicycle" and the corresponding number of selected objects and images. We can see that 7053 objects (9.79%) are bicycles and that 3399 images (41.28%) have at least one bicycle. This feature enhances data analysis by maintaining context across data dimensions.

analysis of time series data, simplifying essential tasks such as regression. Furthermore, its functionality seamlessly extends to computer vision applications. Indeed, DebiAI provides tailored visual support for each stage of the process, enhancing models in tasks such as object detection and image classification.

In the following two sections, we present the use of multiple widgets in DebiAI for various use cases and provide an overview of a use case related to 2D object detection.

A. DebiAI Visual Functionalities

As described in Section III-A, DebiAI gives the ability to visualize and create interactive dashboards. Moreover, it can visualize various data types such as time series, point clouds and tabular data and display computed attributes of images. However, for images viewing, it can establish links with external tools. In this section, we review a set of graphics implemented on different datasets with different data types. We also illustrate the main filtering features proposed by DebiAI. Four graphical visualizations are presented by exploring the parallel coordinates, the data distribution, the points plots and the time series widgets enhanced with interactive options. The following figures (8 to 11) described below are illustrating the main function with adapted filter. However, to be able to read and distinguish between selected filters, the dedicated website <https://demo.debiai.fr/#/> offers the opportunity to reproduce the same figures.

Fig. 8 illustrates a visualization of a dataset by using a parallel coordinates and the possibility to filter directly a set of variables. Another graphic visualization to analyze data distribution variables with the possibility of grouping by other variables is shown in Fig. 9. The third visualization selected from DebiAI is the possibility to apply statistical measures.

Fig. 10 captures a data cloud visualized with its primary statistical measures; an envelope of min and max of the data,

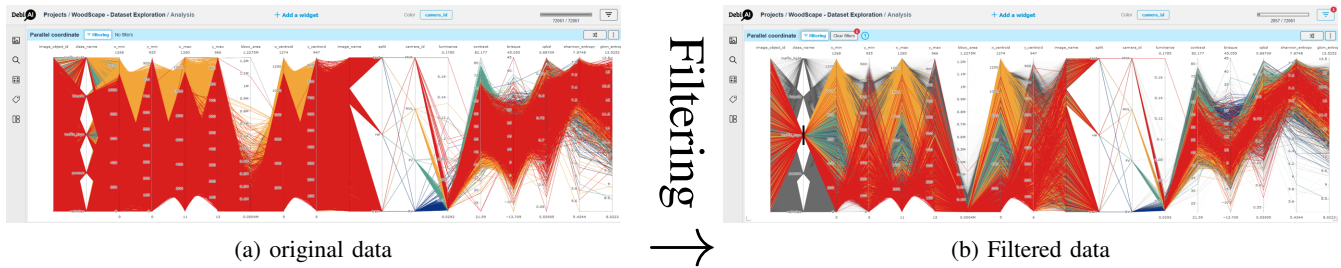


Fig. 8: parallel coordinates widget. (a) represents the original data uploaded and (b) represents the same widget by selecting a subset of variables interactively.

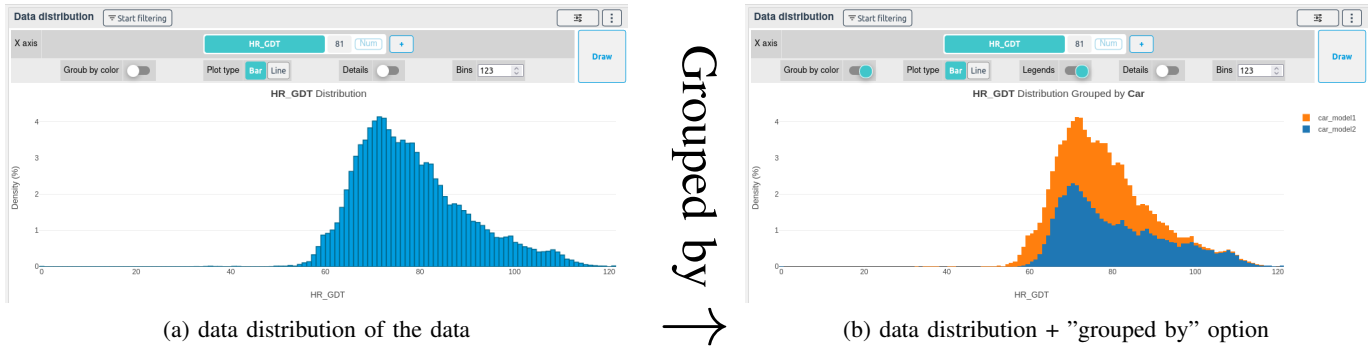


Fig. 9: illustration of data distribution by adding the option of "grouped by". (a) represents an example of data and (b) represents the same data grouped by another variable with two different colors.

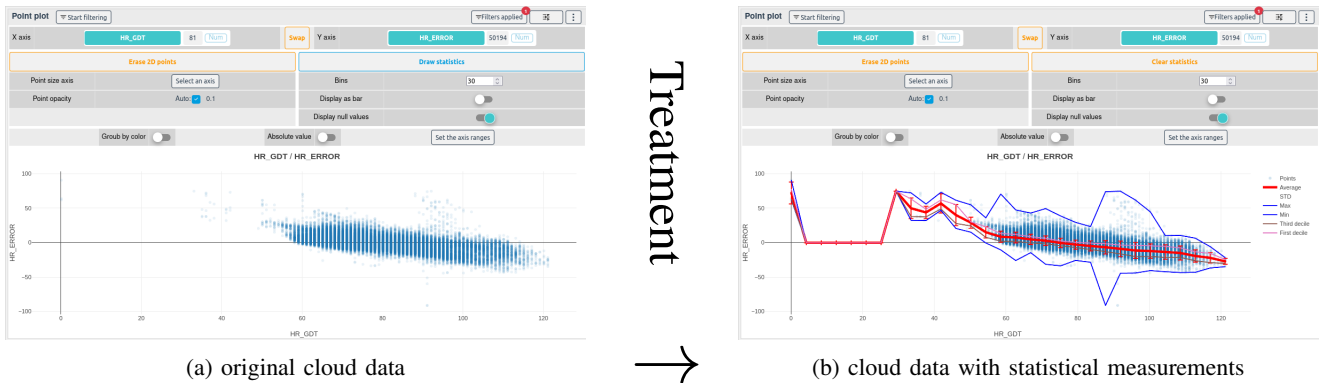


Fig. 10: Statistical treatment for an example of cloud data. (a) represents the original data and (b) represents the data by adding a set of statistical measures. Here illustrated measures are: mean, standard deviation, min and max and deciles

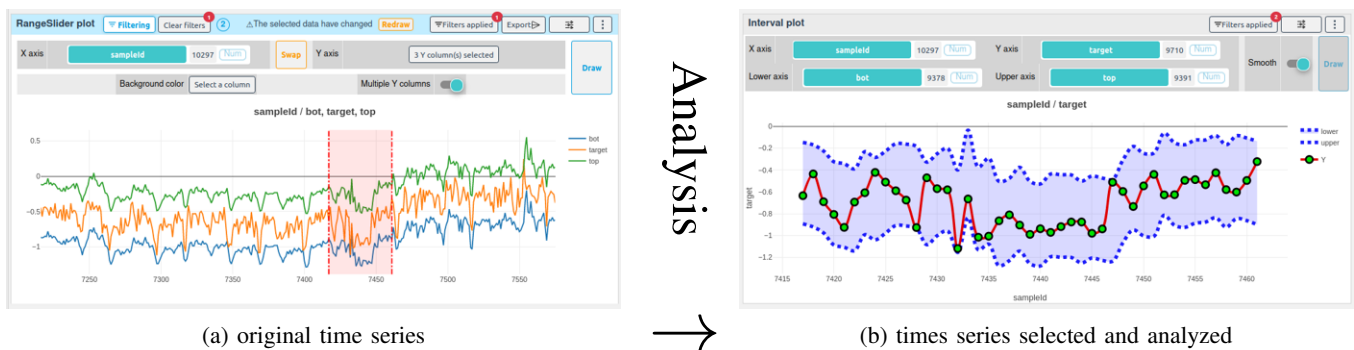


Fig. 11: Statistical analysis applied on times series example. (a) represents the original time series (b) represents the times series filtered and analyzed by adding a set of statistical measures

the average, the confidence interval $\pm\sigma$, where σ represents the standard deviation and also two deciles of the data. Fig. 11 visually encapsulates the two distinct stages in the statistical analysis of a time series. Initially, Fig. 11a displays the original data over an extended period. Subsequently, in the second stage, Fig. 11b illustrates the time series after a more detailed analysis and filtering, focusing on a shorter timeframe. Among the statistical measures, a noteworthy one to explore when analyzing dataset's variables is examining their correlation, a task effortlessly accomplished using the DebiAI's correlation matrix widget.

Fig. 12 displays the Pearson and Spearman correlation matrices for a set of a dataset's image quality metrics. We observe that contrast is positively correlated to Shannon entropy in both correlation matrices, which suggests a monotonic affine relationship between contrast and Shannon entropy attributes.

It is worth noticing that DebiAI's is part of a broader environment and can host tools such as Data Quality Metrics (DQM) library [29]. A visualization of this integration is provided in Fig. 13.



Fig. 12: DebiAI's visualization of correlation Matrices of the WoodScape Dataset under two forms: Pearson (left) and Spearman (right).

B. Images Dataset: 2D Object Detection

In order to illustrate DebiAI's functionalities on images datasets, we conducted some experiments using the WoodScape dataset [28], [30]. WoodScape is a public dataset containing more than 100k images of urban scenes captured using fish-eye cameras for automotive driving tasks from three distinct geographical locations: USA, Europe and China. The images are provided by four cameras with different angles of view (front, rear, middle right and middle left) with 360° coverage and have annotations for a diverse set of computer vision tasks. In addition to different calibration and vehicle information, The dataset provides annotations of 10k images for nine tasks; for instance, 2D object detection, semantic, instance and motion segmentation.

In the scope of this work, we focused on the 2D object detection task for five classes: vehicle, person, bicycle, traffic light and traffic sign. Our selection from the dataset was split into three chunks of 6 : 1 : 3, namely, train, validation and test, respectively. The study process is divided into two main steps: i) data comprehension and ii) results exploration.

The first step aims to obtain a comprehensive overview of the data distribution, understand its scope and how it can be effectively used in a ML process. This comprehension is crucial, as it helps transform an industrial problem into a ML task and establish the appropriate process for models training and results validation. Fig. 14 displays the train set's final distribution grouped by cameras IDs using DebiAI's Data Distribution widget. By applying the same configuration to display the distribution of each of the three sets (train, validation and test), we observed a similar distribution among the three of them. Nevertheless, the figure highlights an immense imbalance among the distribution of the five classes, which is essential for: i) using the appropriate adaptive training techniques, for instance, a weight sampler, and ii) taking into consideration this imbalance when interpreting the models outputs to avoid biased and skewed conclusions.

In the second stage, we used DebiAI to analyze the results of our models applied on the WoodScape test set and put them back into the context of the dataset and its features. This approach ensures an accurate interpretation of the models' outputs and provides potential improvements directions. In our experiments, we used two versions of YOLO-based architectures, specifically YOLOv5 and YOLOv8. The first model, a YOLOv5 with COCO2017 [31] weights. The second model is a YOLOv5 and the third one is a YOLOv8 both trained on WoodScape dataset. Fig. 15 illustrates the relationship between the precision and the recall of each model using the *night stars plot* widget, which helps to navigate the trade-off between the two depending on the context of the task, for instance, are we prioritizing the detection quality over the quantity and vice-versa.

Fig. 16 shows the distribution of the F1-score of each model grouped by camera IDs, where we can easily spot the gap in performance between the three models: having the two models trained on WoodScape dataset showing higher scores compared to the one pre-trained on COCO2017 dataset, which is expected giving the discrepancy between the two datasets. We can also notice that the YOLOv5 trained on the WoodScape train set has better score on the data coming from the front and rear view cameras (FV and RV) of the vehicle while the YOLOv8 also trained on WoodScape shows a greater score on the middle view cameras (left and right) data. This first observations led to further investigations using DebiAI in an attempt to understand the models' outputs; you can check our tutorial on our website for more details, where a complete tutorial is given in the following link: <https://debiai.irt-systemx.fr/tutorials/woodscapeTutorial/>.

C. Time-series Dataset: Anomaly detection in time-series

We conduct experiments concerning the usage of DebiAI on anomaly detection on time-series applied to Server Machine Dataset (SMD) [32]. This dataset is commonly used in benchmark dataset for anomaly detection in time-series. It deals with the monitoring of the performance and health of server machines in data centers. This dataset has been collected by measuring 38 several key performance indicators (KPI) and

Use an algorithm (i)

+ Add a new Algo Provider Refresh Close

Integrated Algo-provider <small>/app/algo-provider</small>	✓ Available	2	Delete
DQM <small>http://172.17.0.3:3020</small>	✓ Available	7	Delete

chisquare test 1.0.0

chi-square test

Analysis of data distribution using a chi-square test for goodness of fit. It supports normal and uniform distributions.

Use algorithm

Created by Confiance.ai

Inputs:

data Array of numbers

Set of data

bins number

Number of bins

distribution string (i)

distribution name

Outputs:

p-value number

The p-value from the chi-square test

intervals frequencies Array of numbers

The data containing observed and expected frequencies

Fig. 13: Illustration of Algo-provider services in DebiAI. Integration of a new Algo-provider named DQM developed in Confiance.ai Program. The new algorithms can use directly filtered/selected dataset from the project.

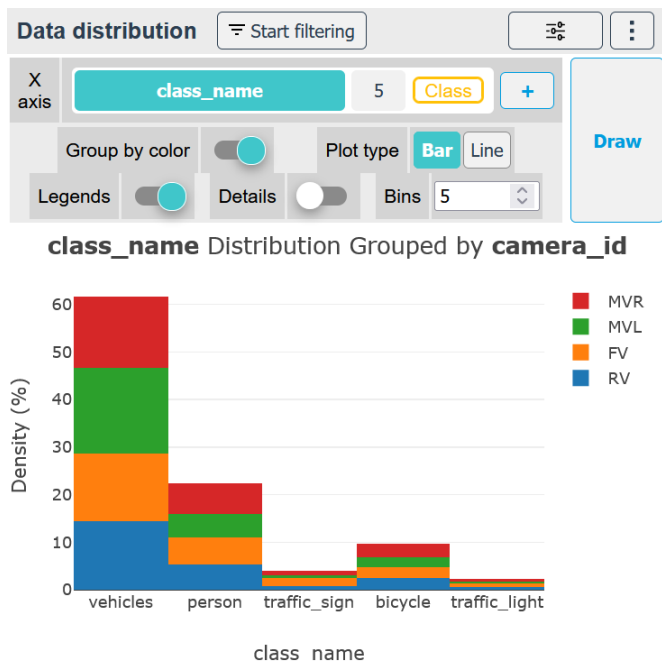


Fig. 14: DebiAI's visualization of WoodScape train's set distribution grouped by Camera ID: MVR (Middle View Right), MVL (Middle View Left), FV(Front View) and RV(Rear View)

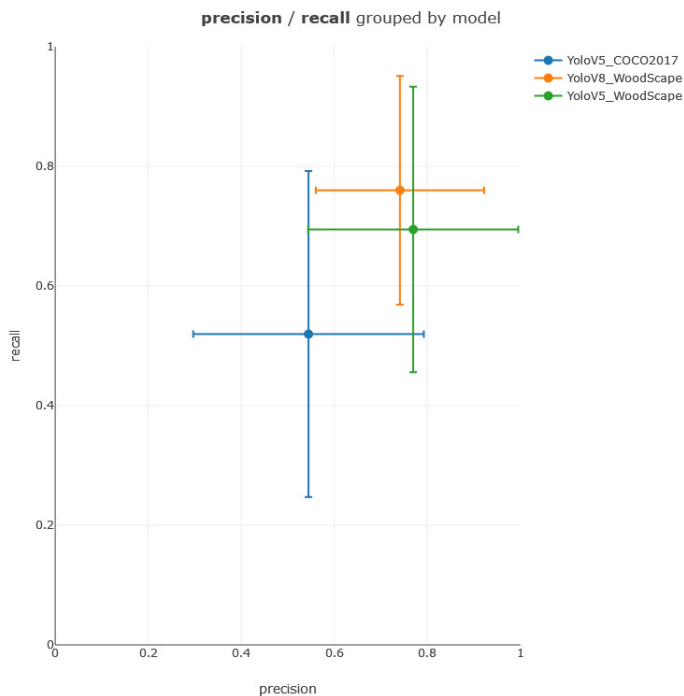


Fig. 15: DebiAI's visualisation of the Precision - Recall relationship of the three models: YOLOv5 on COCO2017, YOLOv5 on WoodScape and YOLOv8 on WoodScape.

anomalies occurrences from 28 different machines during 5 week. KPIs are indicators such as CPU usage, memory usage, system loaded.

Fig. 17 illustrates the capacity of DebiAI on producing clear and insightful SMD representations of 5 machines' first KPI. DebiAI's RangeSlider functionality renders two plots on

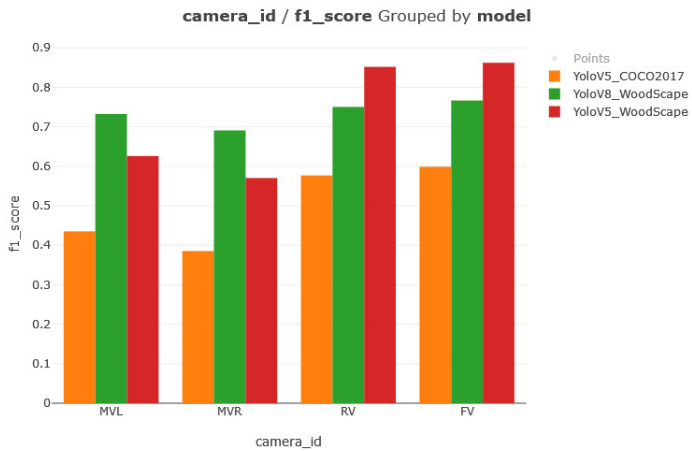


Fig. 16: DebiAI’s visualisation of the F1-score results by Camera ID grouped by Models.

a desired range of timestamps, this plot allowing a detailed exploration of zoomed areas of interest while the below plot gives reference on the global time-series visualisation to better situate the user’s zoomed explorations by giving them a global picture reference. Both plots present the same annotations: 5 different time-series, one for each machine’s first KPI.

We trained a Topological Anomaly Detector (TAD) [33] algorithm on a time series from a single machine in the dataset. Developed by IRT SystemX and available through the TADkit (Time-series Anomaly Detection kit) library, the TAD algorithm is an unsupervised method from TDAAD (Topological Data Analysis for Anomaly Detection) a TADkit’s component designed to detect anomalies in time series data. It assigns an anomaly score to each timestamp, with anomalies identified based on a predefined threshold. To evaluate the algorithm’s performance, we compared its predictions against the ground truth anomaly labels in the dataset using DebiAI.

DebiAI RangeSlider plot shown in Fig. 18 visualizes the performance of our model by displaying a time series curve overlaid on a color-coded background that distinguishes between different prediction outcomes. The background colors represent the four possible classification results: True Positives (TP) are shown in red, indicating correctly detected anomalies; False Positives (FP) appear in yellow, representing false alarms where normal behavior was misclassified as an anomaly; True Negatives (TN) are colored in blue, marking correctly identified normal behavior; and False Negatives (FN) are highlighted in green, showing missed anomalies. The curve represents the actual data points, and the color-coded sections help illustrate where the model performed correctly and where it misclassified, giving an intuitive understanding of the prediction results over time.

VI. EVALUATION ANALYSIS AND USAGE RECOMMENDATIONS

As presented above, DebiAI offers both a Python module and a web-based Data-provider as alternatives for adding data

into a project environment. In this section, we assess these services in terms of scalability, exhibit a specific trade-off between these two approaches and propose user recommendations on how to use them.

To benchmark the performance of DebiAI’s two services above, we evaluated the insertion time provided by the application, whether it’s the DebiAI Python Module or the Web Data Provider, by inserting randomly generated CSV files with varying sizes : 1 000, 10 000, 100 000 and 1 000 000 samples.

TABLE II: Time, in seconds (s), of data insertion for variable samples sizes along DebiAI’s Web-based Data Provider and Python Module services.

Samples	Web Data-Provider	Python Module
1000	0.1s	1s
10 000	0.4s	8s
100 000	3s	62s
1 000 000	33s	465s

The evaluation of DebiAI’s services, shows contrasting performance behaviors as dataset sizes increase. On the one hand, we observe a logarithmic trend on the side of DebiAI’s Web Data-Provider, highlighting Web Data-Provider’s scalability and consistent efficiency, even when handling large datasets. On the other hand, the DebiAI’s Python Module demonstrates an exponential growth in insertion time, indicating a sharp increase in computational demands as the dataset size increases. While Python Module handles smaller datasets reasonably well, the exponential rise in insertion time for larger datasets, such as 1 000 000 samples, reveals significant scalability limitations for this alternative as can be easily seen in Table II.

This sheds evidence on a trade-off between accessibility and scalability arising on both DebiAI’s services. We suggest to use DebiAI’s Python Module as a quick-start approach for usage, and while the amount of data increases in the project environment, switching to the Web Data-Provider service included in DebiAI.

VII. LIMITATIONS

DebiAI is currently in its beta phase, which introduces several limitations. Not all functionalities have been fully implemented, such as complete persistence when navigating away from and returning to the analysis page, which may disrupt the user workflow. Additionally, as demonstrated in our benchmarking (Section VI), the platform’s performance can degrade when handling large datasets exceeding 1 million samples, indicating a need for further optimizations. Users may also encounter occasional bugs or incomplete features, as the tool continues to evolve with new additions and improvements planned for future releases. Despite these limitations, DebiAI is actively used within our company, where it continues to add significant value to our machine learning workflows and projects.

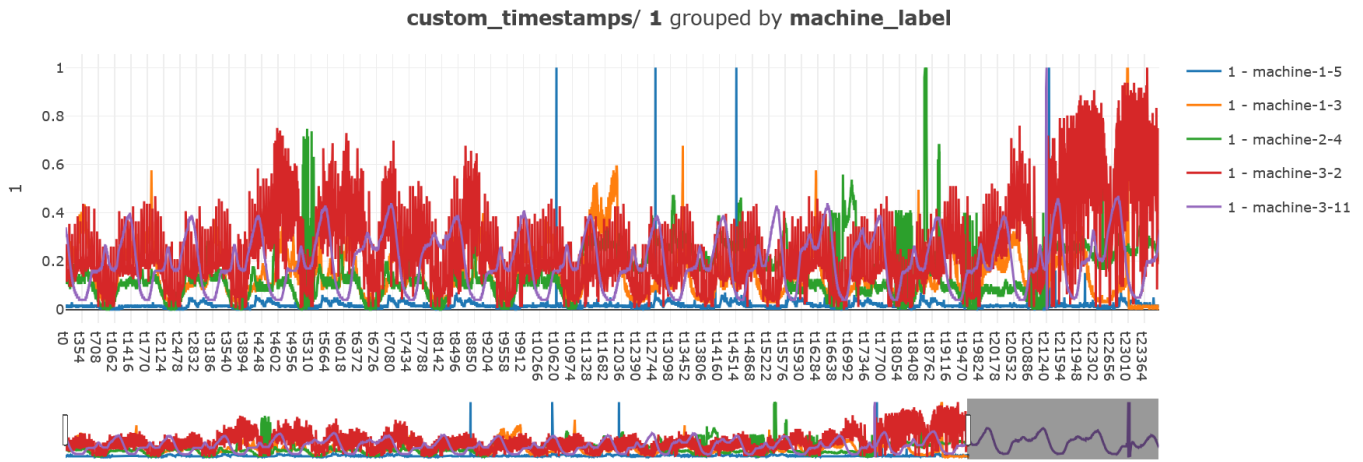


Fig. 17: DebiAI’s RangeSlider provides a comprehensive representation of each of the 5 chosen machines’ first KPI in the SMD dataset.

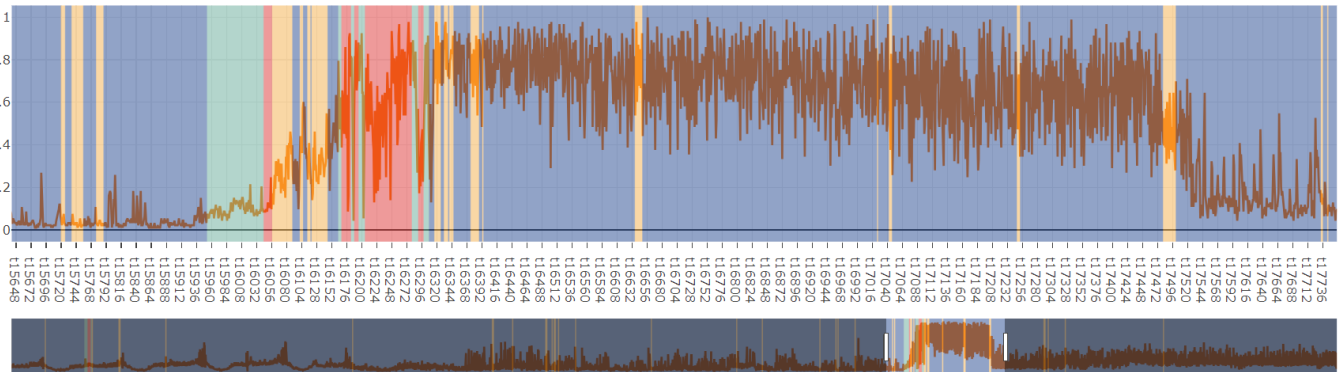


Fig. 18: DebiAI’s RangeSlider applied on a single machine timeseries with anomaly detection algorithm performances compared to dataset ground truth using background plot colors on first KPI. TP: red, FP: yellow, TN: blue, FN: green

VIII. CONCLUSIONS AND PERSPECTIVES

In this paper, we introduced DebiAI, a versatile web-based visual analytics tool that enhances data preparation process, quality assessment, model results analysis and comparison in ML projects. Its adaptability to various use cases and user-friendliness make it a valuable asset contributing to the trustworthiness in AI. For instance, we illustrated its application in a use case of 2D object detection task for driving assistance and server machine sensors analysis for machine’s performances monitoring. As Machine Learning evolves, DebiAI can play a pivotal role in ensuring reliable and interpretable ML outcomes, solidifying its relevance in the field. Compared to our previous version, we simplified data and algorithms insertion as well as installation process with debiai-gui and docker. In DebiAI’s outlook, the priorities are to enhance interoperability with the learning process to retrieve and analyze data from each cycle. The concepts of robustness and explainability are also tied to model’s quality. Therefore, incorporating these metrics into the process is critical for overall trust. We aim to make it easier for users to integrate

and run their own algorithms, be it custom or from a specific library such as DQM and TADkit libraries.

ACKNOWLEDGMENTS

This work has been supported by the French Government under the “France 2030” program, as part of the SystemX Technological Research Institute. This work was conducted as part of the Confiance.ai program, which aims to develop innovative solutions for enhancing the reliability and trustworthiness of AI-based industrial systems. The team, especially T. M. & F. A., would like to thank Rémi Boyer, from IRT SystemX, for his feedback while testing DebiAI’s different features and tutorials, and Jaime De Oliveira along with his team at Thales for their encouraging support and suggestions along the construction of this project.

REFERENCES

[1] T. Mansion, R. Braud, A. Amrani, S. Chaouche, F. Adjed, and L. Cantat, “DebiAI: Open-Source Toolkit for Data Analysis, Visualisation and Evaluation in Machine Learning,” in *ICAS 2024*, 2024.

- [2] F. Sperrle, M. El-Assady, G. Guo, R. Borgo, D. H. Chau, A. Endert, and D. Keim, "A Survey of Human-Centered Evaluations in Human-Centered Machine Learning," *Computer Graphics Forum*, vol. 40, no. 3, pp. 543–568, Jun. 2021, <https://doi.org/10.1111/cgf.14329>, last accessed on 2024-11-20.
- [3] T. Kaluarachchi, A. Reis, and S. Nanayakkara, "A review of Recent Deep Learning Approaches in Human-Centered Machine Learning," *Sensors*, vol. 21, no. 7, p. 2514, 2021.
- [4] T. Capel and M. Brereton, "What is human-centered about human-centered AI? a map of the research landscape," in *Proceedings of the 2023 CHI conference on human factors in computing systems*, 2023, pp. 1–23.
- [5] D. Sacha, M. Sedlmair, L. Zhang, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, and D. A. Keim, "What You See is What You Can Change: Human-centered Machine Learning by Interactive Visualization," *Neurocomputing*, vol. 268, pp. 164–175, 2017.
- [6] Confiace.ai, "Confiace.ai Program," <https://www.confiance.ai>, 2024, online; accessed 19-February-2024.
- [7] M. Adedjouma, C. Alix, L. Cantat, E. Jenn, J. Mattioli, B. Robert, F. Tschirhart, and J.-L. Voirin, "Engineering Dependable AI Systems," in *17th Annual System of Systems Engineering Conference (SOSE)*. Rochester, United States: IEEE, Jun. 2022, <https://hal.science/hal-03700300>, last accessed on 2024-11-29.
- [8] M. Juliette, L. R. Xavier, B. Bertrand, C. Loic, T. Fabien, R. Boris, G. Rodolphe, and N. Yves, "AI Engineering to Deploy Reliable AI in Industry," *AI4I*, vol. 25, no. 1, 2023.
- [9] Taylor, "TAYLOR Network Program," <https://tailor-network.eu/>, 2024, online; accessed 19-February-2024.
- [10] J. Yuan, C. Chen, W. Yang, M. Liu, J. Xia, and S. Liu, "A Survey of Visual Analytics Techniques for Machine Learning," *Computational Visual Media*, vol. 7, no. 1, pp. 3–36, Mar. 2021, <https://doi.org/10.1007/s41095-020-0191-7>, last accessed on 2024-11-29.
- [11] J. Wang, S. Liu, and W. Zhang, "Visual Analytics For Machine Learning: A Data Perspective Survey," *CoRR*, vol. abs/2307.07712, 2023, <https://doi.org/10.48550/arXiv.2307.07712>, last accessed on 2024-11-29.
- [12] F. Hohman, K. Wongsuphasawat, M. B. Kery, and K. Patel, "Understanding and Visualizing Data Iteration in Machine Learning," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–13.
- [13] J. Françoise, B. Caramiaux, and T. Sanchez, "Marcelle: Composing Interactive Machine Learning Workflows and Interfaces," in *The 34th Annual ACM symposium on user interface software and technology*, 2021, pp. 39–53.
- [14] S. Kandel, R. Parikh, A. Paepcke, J. M. Hellerstein, and J. Heer, "Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*, 2012, pp. 547–554.
- [15] M. Chakraborty and A. P. Kundan, "Grafana," in *Monitoring Cloud-Native Applications: Lead Agile Operations Confidently Using Open Source Software*. Springer, 2021, pp. 187–240.
- [16] J. Turnbull, *Monitoring with Prometheus*. Turnbull Press, 2018.
- [17] K. Ahmad and M. Ansari, "Hands-on influxdb," in *NoSQL*. Chapman and Hall/CRC, 2017, pp. 341–354.
- [18] S.-E. Hong, J. Moon, and J. Lee, "Data Acquisition and Visualization for AI/ML-based Radio Resource Management Optimization in the ns-0-RAN Framework," in *2024 Fifteenth International Conference on Ubiquitous and Future Networks (ICUFN)*. IEEE, 2024, pp. 476–478.
- [19] E. G. Rani and D. Chetana, "Using GitHub and Grafana Tools: Data Visualization (DATA VIZ) in Big Data," in *Computer Vision and Robotics: Proceedings of CVR 2022*. Springer, 2023, pp. 477–491.
- [20] A. Vulpe, C. Dobrin, A. Stefan, and A. Caranica, "AI/ML-based Real-Time Classification of Software Defined Networking TSraffic," in *Proceedings of the 18th International Conference on Availability, Reliability and Security*, 2023, pp. 1–7.
- [21] M. Khorasani, M. Abdou, and J. H. Fernández, "Web Application Development with Streamlit," *Software Development*, pp. 498–507, 2022.
- [22] E. Haedecke, M. Mock, and M. Akila, "ScrutinAI: A Visual Analytics Tool Supporting Semantic Assessments of Object Detection Models," *Computers & Graphics*, vol. 114, pp. 265–275, 2023, <https://www.sciencedirect.com/science/article/pii/S009784932300105X>, last accessed on 2024-11-20.
- [23] J. Zhang, Y. Wang, P. Molino, L. Li, and D. S. Ebert, "Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 364–373, 2019.
- [24] C. Chen, Y. Guo, F. Tian, S. Liu, W. Yang, Z. Wang, J. Wu, H. Su, H. Pfister, and S. Liu, "A Unified Interactive Model Evaluation for Classification, Object Detection, and Instance Segmentation in Computer Vision," *IEEE Transactions on Visualization and Computer Graphics*, 2023.
- [25] S. Amershi, M. Chickering, S. Drucker, B. Lee, P. Simard, and J. Suh, "ModelTracker: Redesigning Performance Analysis Tools for Machine Learning," in *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2015)*. ACM - Association for Computing Machinery, Apr. 2015, <m/en-us/research/publication/modeltracker-redesigning-performance-analysis-tools-for-machine-learning>, last accessed on 2024-11-29.
- [26] D. Ren, S. Amershi, B. Lee, J. Suh, and J. D. Williams, "Squares: Supporting Interactive Performance Analysis for Multiclass Classifiers," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 01, pp. 61–70, jan 2017.
- [27] C. Wang, W. Shi, J. Zhang, W. Wang, H. Pan, and F. Feng, "Debias Can be Unreliable: Mitigating Bias Issue in Evaluating Debiasing Recommendation," *arXiv preprint arXiv:2409.04810*, 2024.
- [28] S. Yogamani, C. Hughes, J. Horgan, G. Sistu, P. Varley, D. O'Dea, M. Uricar, S. Milz, M. Simon, K. Amende, C. Witt, H. Rashed, S. Chen-nupati, S. Nayak, S. Mansoor, X. Perroton, and P. Perez, "WoodScape: A Multi-Task, Multi-Camera Fisheye Dataset for Autonomous Driving," 2021.
- [29] S. Chaouche, Y. Randon, F. Adjed, N. Boudjani, and M. Ibn Khedher, "DQM: Data Quality Metrics for AI Components in the Industry," in *The symposium on AI Trustworthiness and Risk Assessment for Challenged Contexts (ATRACC)*, 2024, pp. 1–8.
- [30] S. Yogamani, C. Hughes, J. Horgan, G. Sistu, P. Varley, D. O'Dea, M. Uricár, S. Milz, M. Simon, K. Amende *et al.*, "Woodscape: A Multi-Task, Multi-Camera Fisheye Dataset for Autonomous Driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9308–9318.
- [31] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," 2015.
- [32] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei, "Robust Anomaly Detection for Multivariate Time Series Through Stochastic Recurrent Neural Network," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2828–2837.
- [33] R. Martin. (2024) Topological Data Analysis for Anomaly Detection. https://confianceai.pages.irt-systemx.fr/ec_5/ec5_as3/tdaad/tech_docs.html, last accessed on 2014-09-15.

Development of Children's Crossing Skills in Urban Area: Visual Exploration and Mental Representation about Hazards

Jordan Solt
Lorraine University, 2LPN
Luxembourg-ville
Email: solt.jordan@gmail.com

Aurelie Mailloux
Reims hospital, URCA
Reims, France
Email: aurelie.mailloux@univ-reims.fr

Jerome Dinet
Lorraine University and Chair BEHAVIOUR
Nancy, France
Email: jerome.dinet@univ-lorraine.fr

Samuel Ferreira Da Silva
Lorraine University and Chair BEHAVIOUR
Metz, France
Email: samuel.ferreira-da-silva@univ-lorraine.fr

Muneo Kitajima
Nagaoka University of Technology
Nagaoka, Niigata, Japan
Email: mkitajima@kjs.nagaokaut.ac.jp

Gaelle Nicolas
Lorraine University and Chair BEHAVIOUR
Metz, France
Email: gaelle.nicolas@univ-lorraine.fr

Abstract—Pedestrian trauma represents a significant proportion of all road traumas, young pedestrian being over-represented in all these road traumas. From a cognitive point of view, road crossing ability is a high and complex mental activity because the individual has to process dynamic and complex information extracted from his/her surrounding environment, to make a decision (i.e., where and how to cross), and safe pedestrians must possess and utilize advanced cognitive skills. More precisely, there are two major problems for young pedestrians to make the decision about when and where it is safe to cross the street: gap selection and assessment of inter-vehicular gap. A first study conducted with forty children aged 3-10 years and twenty-two adults has been conducted to investigate the impact of one individual factor (Age) and one environmental factor (Traffic density) on decision making (i.e., “to cross” or “not to cross a street”), time spent to make decision (in milliseconds) and on visual exploration using eye-tracking techniques of urban scenes displayed on a computerized screen. Main results showed that (i) Traffic density has a significant impact on performance and visual exploration, (ii) Age has a significant impact on time spent to make decision and visual exploration and (iii) there is an interaction between Age and Traffic density. A second study, based on drawings performed by 125 young pedestrian provided relevant qualitative and subjective data to complete data issued from the first experiment based on eye-tracking technique. The children's drawings are more elaborate and richer (in terms of the number of elements represented) as the real accident risk in their environment increases. Not only does the number of elements in the drawings increase quantitatively, but these elements are more closely linked to each other, to the point where they form a real scene. The mental representation becomes more and more precise, and the egocentric perspective is dominant, this result being consistent with the Piagetian approach to the development of intelligence for children.

Keywords—Child; Pedestrian; Visual exploration; Risk; Hazard; Eye-Tracking; Drawing; Perceived risks

I. INTRODUCTION

Because pedestrian trauma represents a significant proportion of all road traumas, and because young pedestrian being over-represented in all these road traumas [1], the safety of child pedestrians is of concern, given that a sizable proportion of pedestrians killed and seriously injured involve children and the special value society places on its youth [2][3]. Although there is no consensus definition, most data sets identify pedestrians as those walking, running, standing, or lying on a road, right of way, or parking lot [4][5] and people on bicycles, skateboards, and other non-powered conveyances are not considered pedestrians for the purposes of this report. Pedestrian injury data are often stratified into “traffic” data, which include those injured or killed while on the roadway, and “non-traffic” data involving those struck on driveways or private lots, an important subset of the injured among younger children

The structure of this paper is the following:

- In Section I, context and challenges related to accidents with young pedestrians and factors influencing children's crossing skills are presented;
- In Section II, two studies conducted with our participants are described: The first study is based on eye-tracking data issued from an experiment conducted with children aged 3-10 years-old to investigate the impact of several specific factors on gaze exploration and decision-making; The second study is based on analyses of drawings performed by young pedestrians. In other words, the first data are quantitative while the data extracted from drawings are more subjective;
- Finally, in Section III, theoretical and methodological implications related to the changes in visual strategy occurring around the age of 7-8 years are discussed.

A. Context

Around the world, the number of pedestrians killed increases. Young pedestrians are particularly concerned by these accidents: According to the official data issued from the Traffic Safety Facts, on average, three children were killed and an estimated 502 children were injured every day in the U.S. in traffic crashes. In 2019 and 2020, there were respectively 181 and 177 children killed in pedestrian accidents. Most were toddlers (between the ages of 1-3) and young children (4-7). In fact, an estimated 1 in 5 children killed in car accidents were pedestrians, i.e., just walking on the sidewalk or crossing a street whatever the country [6][7].

At ages 6-10 years, children are at highest risk of pedestrian collision, most likely due to the beginning of independent unsupervised travel at a time when their road strategies, skills and understanding are not yet fully developed [8]. Whatever the country, research suggests that children between the ages of 6 to 10 are at highest risk of death and injury, with an estimated minimum four times the risk of collision compared to adult pedestrians [9]. Until the age of 6-7 years, children are under active adult supervision, i.e., parents hold their child's hand when crossing roads together.

Even if every year many pedestrians are injured or killed in traffic accidents in rural parts of the country [10], pedestrian safety is being considered as a serious traffic safety problem in urban and suburban settings [11][12]. Thus, children more than adults, are at risk as pedestrians, often due to their own actions and behaviors. So the question is: "Why do young pedestrians not adopt safety behaviors specially during street crossing?"

B. Factors Influencing Children's Crossing Skills and Gap Selection

From a cognitive point of view, road crossing ability is a high and complex mental activity because the individual has to process dynamic and complex information from his/her surrounding environment, to make a decision (i.e., where and how to cross). Safe pedestrians must possess and utilize advanced cognitive skills [13][14]. Crossing decisions include whether or not to enter the roadway, the place to cross, the path to take, how fast to travel, and how the driver might react. A sound decision on whether to enter the roadway should be based upon recall (experience) and monitoring of the traffic detected, including the distance, speed, and anticipated direction of vehicles and the opportunities provided by various gaps in traffic [15]. The time that has elapsed while making the decision also needs to be incorporated. Successful crossing performance also requires reliable estimation of the pedestrian's walking speed, peak capabilities, and distance to the other side of the road or a traffic island. Integrating all these aspects is difficult for the child, especially one inexperienced in traffic, and result in a longer decision making time: In fact, a 5 year old child requires about twice as long to reach a pedestrian decision as an adult, and this leaves even less time to execute an imperfectly planned crossing [13][14][16].

A vast amount of research suggests that children's development of cognitive skills is significantly related to increased pedestrian safety and that relevant skills improve as children get older [17][18][19]. Of course, it is not a single cognitive skill that influences safety. Instead, it is the combined development of a number of different cognitive processes that are linked to safe pedestrian behavior. Those processes also overlap with other developing skills, such as perceptual (visual and auditory essentially) and motor abilities.

As children develop, specific pedestrian injury risks change [16][19][20][21][22][23]. More precisely, toddlers (ages 1-2) are most likely to be injured in driveways, where drivers moving backward are unable to see them [24], while adolescents are at risk due to walking at night with poor visibility, walking while intoxicated or walking while distracted by phones [25]. Our paper focuses on children between those two phases, in ages 6 through 12. During this stage of development, most pedestrian injuries occur in mid-block areas, where children enter into the middle of the street and are struck by moving vehicles, or at intersections [26]. As Schwebel and his colleagues said, if some incidents are "dart-out" situations where children enter the street quickly, without thought (i.e., to chase a person, toy, or pet, or to meet someone on the other side of the street), the majority of the incidents/collisions are the result of poor judgment by the child, i.e., s/he believes it to be safe, and enters the street when in fact the situation is not safe [23].

Several studies showed that gap selection and assessment of inter-vehicular gap by young pedestrians are two major problems for young pedestrians to make the decision about when and where it is safe to cross the road [27][28][29]. Inter-vehicular gap is both temporal and spatial because these two parameters are crucial to make the decision in relation to available gaps in the traffic [30]. More precisely, judgement of whether a gap in the traffic is sufficient to safely cross requires the determination of the time gap of the nearest vehicle with the planned crossing line and the assessment of whether this time gap exceeds the time required to cross the road. So, children aged below 10 years have relatively poor skills at reliably setting safe distance gap thresholds, and thus do not consistently make safe crossing decisions [31][32][33][34][35][36].

But, very few authors concentrated on visual exploration of young pedestrians during crossing activity. For instance, Whitebread and her colleagues examined the relationships between pedestrian skills and visual search strategies for young pedestrians [37]. According to their findings, major changes in strategy occurred around the age of 7-8 years. This change expressed in the frequency and pattern of looking at different directions, having a sophisticated 'last-minute' checking approach, exhaustive visual search strategy, and the speed of making the crossing decision. In the same way, Tapiro and her colleagues examined children's visual search strategies in hazardous road-crossing situations [33]. A sample of 33 young participants (ages 7-13) and 21 adults observed 18 different road-crossing scenarios in a 180 degrees dome shaped mixed reality simulator. Gaze data were collected

while participants made the crossing decisions. Their results showed that age group, limited field of view, and the presence of moving vehicles affect significantly the way pedestrians allocate their attention in the scene. Therefore, the authors hypothesized that adults tend to spend relatively more time in further peripheral areas of interest than younger pedestrians do. It was also found that the oldest child age group (11-13 years old) demonstrated more resemblance to the adults in their visual scanning strategy, which can indicate a learning process that originates from gaining experience and maturation. Nevertheless, all participants in these previous studies were 7 years old and above. In our experiment, we collected data with eye-tracking from younger pedestrians (3 to 10 years old) to better understand the visual exploration of urban scenes.

II. EXPERIMENT 1: VISUAL EXPLORATION BEFORE TO CROSS A STREET

This experimental study conducted with forty children aged 3-10 years and twenty-two adults was aiming to investigate the impact of one individual factor (Age) and one environmental factor (Traffic density) on decision making (i.e., “to cross” or “not to cross a street”), time spent to make decision (in milliseconds) and on visual exploration of urban scenes displayed on a computerized screen. Eye-tracking technique is used to collect precise data about gaze exploration of each participant.

A. Participants

Sixty-two French participants were recruited to participate in this study. Children are issued from four different age groups: Seven pupils are from Grade 1 (boys, 100 percent; mean age = 3.86 years; SD = 0.37 years), nineteen pupils are recruited from Grade 3 (boys, 56.8 percent; mean age = 6.89 years; SD = 0.31 years), fifteen pupils are recruited from Grade 5 (boys, 60 percent; mean age = 9.87 years; SD = 0.51 years), and twenty-one participants are adults (men, 47.6 percent; mean age = 26.71 years; SD = 8.22 years). All children are issued in the same elementary school located in the mid-town.

All participants are French native speakers and the majority (82.1 percents) lives in urban area. Moreover, even if the majority of adult participants (81 percents) have their driving license, they admit to go to work essentially by using public transportation (61.9 percents) or by walk (38.1 percents). All the children are recruited in the same primary school located in the mid-town. All parents agreed to their children participate. No participant has severe visual impairment and no cognitive impairment. There is no difference between groups according to the visual memory and attention capacities (Table I).

B. Independent and Dependent Variables

In our study, we investigated the impact of one individual factor (Age) and one environmental factor (Traffic density) on three behavioural indicators:

- The decision (i.e., “to cross” versus “not to cross the street”);

- The time spent in milliseconds to make this decision;
- The visual exploration of specific Areas of Interest (AoI) of urban scenes displayed on pictures (Figure 1);

Thus, two independent factors were manipulated, the first one being intra-subject (“Age”, with four modalities: Grade 1, Grade 3, Grade 5, and adults) and the second one being inter-subject (“Traffic density”, with three modalities: Low, Moderate, and High). In other words, our experimental plan was: Participant < Age 4 > * Traffic density 3.

C. Material

Assessment of Cognitive Abilities. Each participant was asked to complete several sub-scales extracted from the Wechsler scales to assess their cognitive abilities. For the youngest participants (Grade 1), “Coding scale” and “Digit span scale” extracted from the WPPSI-V have been used. For the two other groups of children (Grade 3 and Grade 5), they are the same sub-tests used but extracted from the WISC-V. For adults, four sub-scales extracted from the WAIS-V have been used: “Digit span scale”, “Arithmetic scale”, “Coding scale”, and “Symbol scale”. All these sub-scales were chosen because they are very sensitive to the visual memory and attention capacities.

Urban Scenes. Each participant was individually asked to examine different urban scenes displayed on a computerized screen before to make a decision for each urban scene, i.e., “to cross” or “not to cross the street”. Three traffic densities have been used to investigate the impact of this factor on decision-making and visual exploration: Low, Moderate, and High. Figure 1 shows an example for each of these modalities. For each of the traffic density (Low, Moderate, and High), four different urban scenes. These urban scenes were chosen by four judges after they evaluated and categorized a lot of pictures in these three traffic conditions: Low traffic density (“Low”; e.g., one other pedestrian and two vehicles far), moderate traffic density (“Moderate”; e.g., several other pedestrians and different kinds of vehicles), and high traffic density (“High”; e.g., a lot of vehicles near and far).

Each participant was asked to examine 12 different static pictures of urban scenes, the order of presentation being counterbalanced to avoid order effect on responses (i.e., “to cross” or “not to cross the street”). On-line eye-tracking data for each participant were collected during participants examined urban scenes, by using the eye-tracking techniques. The Tobii T120, with a 17 inch monitor integrated, was used to collect visual exploration of urban scenes by our participants.

D. Procedure

The procedure has four distinct and successive steps:

- Training session. First, each participant was invited to seat in front of a computer (Tobii T120, with a 17 inch monitor integrated) and the same instructions are given: (a) different images will appear on the screen, one by one; (b) s/he must to analyse the urban scenes carefully because s/he was asked to decide if s/he crosses or not the street; (c) when s/he made the decision, s/he was asked to say “stop” and s/he can give his/her decision orally.

TABLE I
MEAN (AND STANDARD DEVIATION) OF FIXATION DURATION FOR EACH AGE GROUP, TRAFFIC CONDITION FOR EACH AREAS OF INTERESTS (AOI)

	Mean of Memory Span (SD)	Mean of Processing Speed (SD)
Grade 1 (n = 7)	8.5 (2.7)	9.5 (3.4)
Grade 3 (n = 19)	10.5 (4.5)	12.8 (5.1)
Grade 5 (n = 15)	10.9 (2.9)	10.2 (3.9)
Adult (n = 21)	9.1 (1.9)	10.6 (2.8)

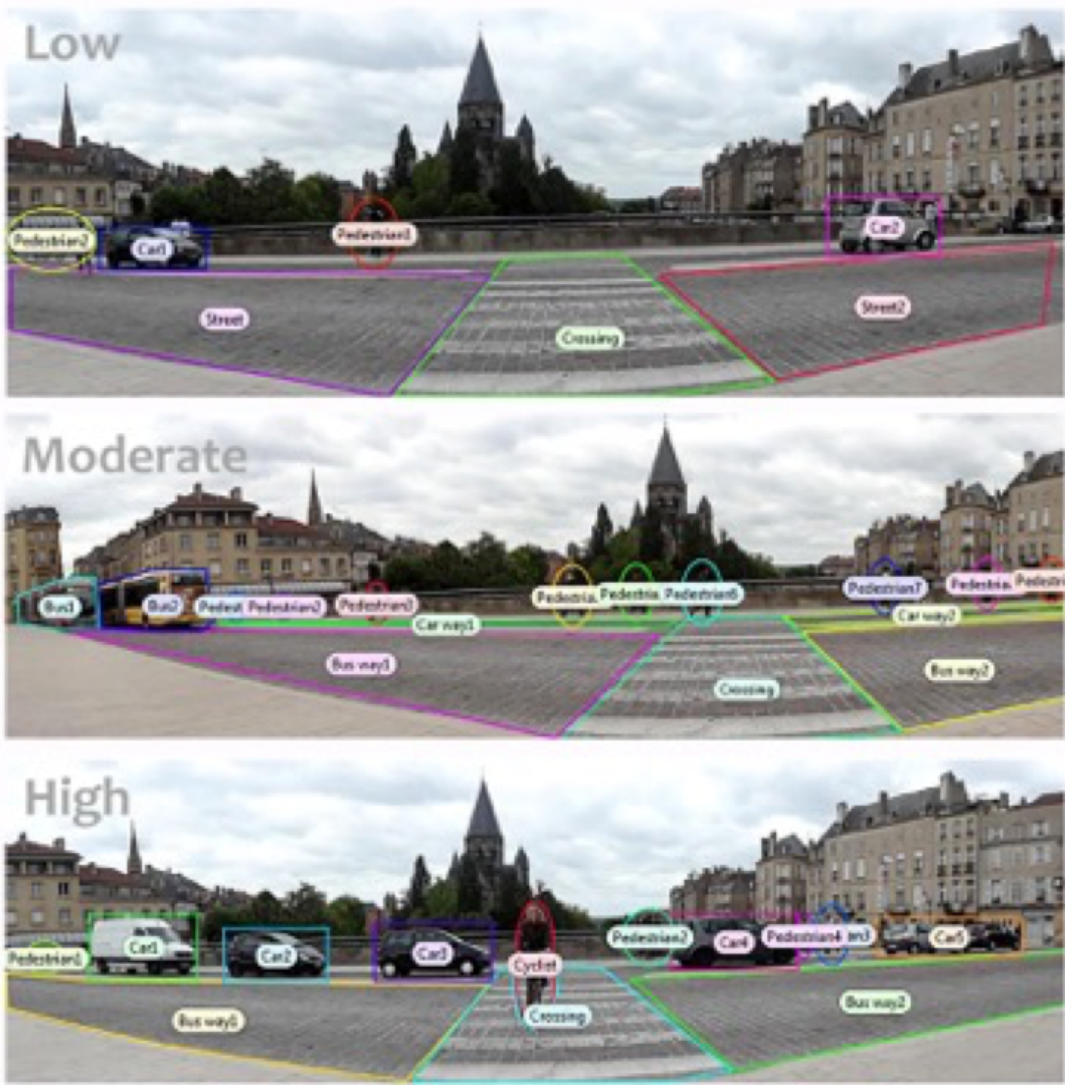


Figure 1. The different Areas of Interest (AoI) in the three Traffic density conditions

Different pictures (not used in the following experiment) are used during a training session;

- Experimental session for visual exploration and decision-making. If the participant has no problem with the procedure and has no question, the experiment can begin with the urban scenes related to the three conditions (Low, Moderate, High);
- Assessment of cognitive abilities. Just after the end of the experimental session, each participant was asked to complete sub-scales extracted from the Wechsler scales to assess their visual memory and attention capacities;
- Length of time the subject is expected to participate
- Researchers ensured that those participating in research will not be caused distress;
- End of the experiment. Finally, each participant was asked to complete a survey to provide some demographic information, is thanked and each child receives a packet of sweets.

Note that for children, the experiment was always conducted in the same quiet room located in the school, dedicated to the experiment. The experimenter was always the same.

E. Design and Data Analysis

First, we examined the impact of our two independent variables (“Group age” and “Traffic condition”) on the one hand, decision (i.e., “I am crossing” or “I am not crossing”), and on the other hand, time spent to make this decision (in milliseconds). So the design of this first part of analyses is the following factorial design: Group age (4) (Grade 1, Grade 3, Grade 5, Adult) X Traffic density (3) (Low, Moderate, High), with “Age group” as between-subjects factor and “Traffic density” as within-subjects factor.

Second, we examined the visual exploration on specific Areas of Interest (AoI) predefined for urban scenes (i.e., “Pedestrians”, “Sidewalk”, “Car”, “Car way”, “Bus”, “Bus way”, and “Crossing”). Figure 1 shows these six different AoI. The design of this second part of analyses is the following factorial design: Age group (4) (Grade 1, Grade 3, Grade 5, Adult) X Traffic density (3) (Low, Moderate, High) X AoI (6) (Pedestrians, Sidewalk, Car, Car way, Bus, Bus way, Crossing) with “Age group” as between-subjects factor and “Traffic density” and “AoI” as within-subjects factors.

F. Ethics

All adults’ participants provided written informed consent for their participation in this study, and all legal parents of children provided the same informed consent. Moreover, the responsible of the school provided also her consent. Before providing the written consent, all adults’ participants, legal parents of children and the director of the school where the research has been conducted received the same information relating to the following points:

- A statement that participation is voluntary and that refusal to participate will not result in any consequences or any loss of benefits that the person is otherwise entitled to receive;

- The precise purpose of the research;
- The procedure and material involved in the research;
- Benefits of the research to society and possibly to the individual human subject;
- Length of time the subject is expected to participate
- Researchers ensured that those participating in research will not be caused distress;
- Subjects’ right to confidentiality and the right to withdraw from the study at any time without any consequences;
- After the research is over, each participant (adults or children) are able to discuss the procedure and the findings with the psychologist.

G. Main Results

The experiment based on eye-tracking techniques aimed to investigate the impact of one individual factor (Age) and one environmental factor (Traffic density) on three behavioural indicators related to competencies of very young pedestrians (aged 3-10 years). Several interesting results have been obtained.

The decision made by each participant (“I cross” versus “I do not cross”) in front of each urban scene has been collected (Table II). For each of the three Traffic density conditions (Low, Moderate, High), statistical analyses revealed only one significant impact of Age group in high traffic condition ($F(3-58) = 2.858$, $p = .045$).

As Table III shows, the time spent to make decision decreased with age. Statistical analyses confirmed that Age group had a significant impact on this time spent to make decision ($F(3-58) = 8.75$, $p < .001$). Time spent to make decision for the youngest participants (Grade 1, Mean = 8829.68) was superior than time spent by all the other participants (Grade 3, $M = 5240.98$, $F(3-58) = 2.934$, $p = .005$; Grade 5, Mean = 4694.68, $F(3-58) = 3.265$, $p < .005$; Adults, Mean = 2797.82, $F(3-58) = 4.996$, $p < .001$). In the same way, time spent to make decision for participants aged to 6-7 years (Grade 3, Mean = 5240.98) was superior than time spent for adults (Mean = 2797.82), the difference being significant ($F(3-58) = 2.789$, $p = .007$). Finally, time spent to make decision for participants in grade 5 (Mean = 4694.68) was superior than time spent adults (Mean = 2797.82), the difference being significant ($F(3-58) = 2.028$, $p = .047$) Traffic condition had also a significant impact on time spent to make decision ($F(2-116) = 7.67$, $p = .001$). As Table II shows, time spent to make decision in low traffic condition (Mean = 4311.31) was inferior than time spent in high traffic condition (Mean = 5278.16), the difference being significant ($F(2-116) = 7.67$, $p = .002$). Finally, there was an interaction between Age group and Traffic condition ($F(6-116) = 2.73$, $p = .016$) on the time spent to make a decision.

There was a global impact of Age group on total fixation duration (Table IV; $F(3-58) = 8.475$, $p < .001$). Specifically, Age group had a significant impact for Low traffic density ($F(3-56) = 2.980$, $p = .039$) and Moderate traffic density ($F(3-56) = 9.422$, $p = .001$) but had no impact on High traffic density ($F(3-50) = 2.695$, $p < .056$):

TABLE II
NUMBER (AND PERCENTAGE) OF PEDESTRIANS CROSSING THE STREET FOR EACH AGE GROUP AND THE THREE TRAFFIC DENSITY CONDITIONS (LOW, MODERATE, HIGH)

	Low	Moderate	High
Grade 1 ($n = 7$)	3 (42.8)	2 (28.5)	1 (14.9)
Grade 2 ($n = 19$)	12 (63.6)	2 (10.5)	0 (-)
Grade 2 ($n = 15$)	10 (66.6)	3 (20)	3 (20)
Adult ($n = 21$)	9 (42.8)	3 (14.2)	5 (23.8)

TABLE III
MEAN (AND STANDARD DEVIATION) OF TIME SPENT TO MAKE DECISION (I.E., "TO CROSS" *versus* "NOT TO CROSS") FOR EACH AGE GROUP AND EACH TRAFFIC DENSITY CONDITION (LOW, MODERATE, HIGH)

	Low	Moderate	High	Mean (SD)
Grade 1 ($n = 7$)	7854 (6399)	7183 (5081)	11450 (10673)	8829 (6812)
Grade 3 ($n = 19$)	4921 (2292)	5337 (1803)	5464 (3140)	5240 (2030)
Grade 5 ($n = 15$)	3804 (1616)	4454 (2175)	5825 (3902)	4694 (1981)
Adult ($n = 21$)	2940 (1708)	2791 (1245)	2661 (1592)	2797 (1345)
Total mean (SD) (N = 62)	4311 (3066)	4469 (2672)	5278 (5025)	-

TABLE IV
MEAN (AND STANDARD DEVIATION) OF TOTAL FIXATION DURATION FOR EACH AGE GROUP AND THE THREE TRAFFIC DENSITY CONDITIONS (LOW, MODERATE, HIGH)

	Low	Moderate	High	Mean (SD)
Grade 1 ($n = 7$)	0.316 (0.08)	0.362 (0.06)	0.408 (0.124)	0.379 (0.09)
Grade 2 ($n = 19$)	0.351 (0.145)	0.357 (0.10)	0.321 (0.14)	0.343 (0.13)
Grade 5 ($n = 15$)	0.303 (0.07)	0.266 (0.06)	0.267 (0.12)	0.265 (0.08)
Adult ($n = 21$)	0.266 (0.05)	0.290 (0.04)	0.278 (0.04)	0.297 (0.06)
Total mean (SD) (N = 62)	0.320 (0.09)	0.311 (0.07)	0.318 (0.11)	-

- For Low traffic density condition, mean fixation duration for children recruited in Grade 1 was higher compared to Adults (respectively, Mean = 0.3614 and Mean = 0.2665; $t(56) = 2.183$, $p = .033$). In the same way, children recruited in Grade 3 have more longer fixation duration compared to Adults (respectively, Mean = 0.3514 and Mean = 0.2665; $t(56) = 2.639$, $p = .011$). Adults had the fastest fixings but that was significant only that in comparison with Grade 1 and Grade 3;
- For Moderate traffic density condition, adults (M = 0.29) had shorter fixation duration compared to Grade 1 (respectively, Mean = 0.29 and Mean = 0.3692; $t(56) = 2.293$, $p = .026$) and compared to Grade 3 (Mean = 0.3579; $t(56) = 2.656$, $p = .01$). Children issued from Grade 5 spent significantly less time to make decision than Grade 1 (Mean = 0.3692; $t(56) = 3.950$, $p = .000$), compared to Grade 3 (Mean = 0.3579) ($t(56) = 4.76$, $p = .000$) and compared to adults (Mean = 0.29) ($t(56) = -2.345$, $p = .023$);
- For High traffic density condition, only one Age group was concerned by significant differences: Children issued from Grade 1 spent significantly more time to make

decision compared to Grade 5, compared to Grade 1 (respectively, Mean = 0.4081 and Mean = 0.2673; $t(50) = 2.521$, $p = .015$) and compared to Adults (Mean = 0.05893; $t(50) = 2.54$, $p = .014$). In other words, in the high traffic density condition, children issued from Grade 1 were the slowest.

As Figure 2 shows, visual fixation duration time was significantly superior for two of the different Areas of Interest (AoI) predefined: the car way ($F(3-43) = 4.191$, $p = .011$) and the crossing ($F(3-55) = 3.891$, $p = .014$).

Moreover, Age group had a significant impact on distribution of fixation time only for these two of the different Areas of Interest (AoI) predefined. Fixation duration time on the car way was superior for Grade 1 compared to Grade 5 (respectively, Mean = 0.3625 and Mean = 0.2444; $t(43) = 2.426$, $p = .02$) and compared to Adults (Mean = 0.2311; $t(43) = 2.626$, $p = .012$). And fixation duration time on the car way was superior for Grade 3 compared to Grade 5 (respectively, Mean = 0.3291 and Mean = 0.2444; $t(43) = 2.329$, $p = .025$) and compared to Adults (Mean = 0.2311; $t(43) = 2.569$, $p = .014$).

The pattern of results was identical for the crossing. Fixation

duration time on the car way was superior for Grade 1 compared to Grade 5 (respectively, Mean = 0.3729 and Mean = 0.2713; $t(55) = 2.3$, $p = .025$) and compared to Adults (Mean = 0.2478; $t(55) = 2.932$, $p = .005$). And fixation duration time on the car way was superior for Grade 3 compared to Adults 5 (respectively, Mean = 0.3248 and Mean = 0.2444; $t(55) = 2.425$, $p = .019$).

Even if there were the only significant differences, some interesting tendencies can be remarked in the Figure 2 for other AoI such as “Pedestrians”, “Cars” and “Bus way”. For these three other AoI, fixation duration time for Adults group is always inferior.

There exist some significant interactions between Age group and Traffic density condition on these fixation duration means for the two main AoI (“Car way” and “Crossing”):

- For Low traffic density condition, fixation duration for younger participants (recruited in Grade 1) was significantly superior than fixation duration for children recruited in Grade 3 specially for “Car way” (respectively, Mean = 0.45 and Mean = 0.263; $t(54) = 2.023$, $p = .048$);
- In the same way, for Moderate traffic density condition, fixation duration for younger participants (recruited in Grade 1) was also significantly superior than fixation duration for children recruited in Grade 3 (respectively, Mean = 0.768 and Mean = 0.438; $t(55) = 3.218$, $p = .002$);
- Finally, for High density traffic condition, fixation duration for children issued from Grade 1 was also superior than fixation duration specially for “cars” (respectively, Mean = 0.430 and Mean = 0.290; $t(56) = 2.019$, $p = .048$). The crossing site was extensively explored by the youngest participants (Grade 1, Mean = 0.442) compared to Adults (Mean = 0.229; $t(50) = 2.413$, $p = .020$) and compared to children recruited in Grade 5 (Mean = 0.374; $t(50) = 2.857$, $p = .006$).

Several interesting results have been obtained in this first experiment. The Traffic density has a significant impact on decision made by all the participants. When there is much information in the urban scene (High traffic condition), less participants decide to cross the street, whatever the Age. Second, the Age has a significant impact on time spent to make decision. The decision-making time decreases when the age increases. This result confirms the fact that the age has a strong impact on decision making in pedestrians’ skills a process which develops and becomes increasingly effective with the age [38][39][40]. Third, there is an interaction between Age and Traffic density: The decision-making time decreases when the age increases specially when there is much information in the urban scene (i.e., High traffic density condition). In a second experiment, we investigated more precisely mental representations of children about hazards in their surrounding physical environments by analyzing drawings performed by these children.

III. EXPERIMENT 2: MENTAL REPRESENTATIONS OF CHILDREN ABOUT THE RISKS IN THE STREETS

Children’s drawings can function as a fascinating window into how kids perceive and represent their world [41]. They are also helpful tools for researchers, because young children sometimes find it easier to communicate with imagery rather than words. Through the process of observing and analyzing the drawings of young children, insights can be gained as to the social/emotional, physical, and intellectual development of each child. Children usually explore the world around them through intellectual, physical and emotional methods for young children; pencil, brush and paper are the best means of conveying their perception of their surrounding environment.

The progression of drawings performed by children can show significant growth and development, as well as determine academic capabilities and skills characteristic of their developmental level [42][43]. For instance, the ability to localize an object with an allocentric or object-centered perspective [44][45], based on reference points external to the body and develops later, when the child becomes conscious of object permanence in space, regardless of the child’s position, is easy to identify in drawings (Figure 5, [45]).

A. Participants

One hundred and twenty-five children agreed to take part in this study. The participants were distributed as follows: 20 children attended a school located in a high-risk zone (7 boys and 13 girls; mean age = 7.6 years, SD = 0.5); 76 children attended a school located in a moderate-risk zone (36 boys and 40 girls; mean age = 7.38 years, SD = 0.67); finally, 29 children attended a school located in a low-risk zone (17 boys and 12 girls; mean age = 7.41 years, SD = 1.05).

The distinction between the three types of zone according to risk (low, moderate, high) was made on the basis of accident data compiled by the Metz Urban Community: The first school is located in the city’s hyper-centre, in a historic district, where road traffic is heavy and previous accidents (i.e., pedestrian-vehicle collisions) confirm that the risk is high. The second establishment is also in the city centre, but in a semi-pedestrian area. There have been few previous accidents, so the road risk is considered moderate. Finally, the last establishment is located on the outskirts of the city centre, in a residential area with little traffic, and the number of previous accidents is virtually nil: the risk is therefore low.

B. Protocol and Design

In order to question children about their perceived risk when they navigate in town, we decided to ask them to draw a picture with the following instructions: “The Little Pedestrian is going to take part in the ‘Pedestrian Challenge’ with you. Can you draw the things s/he needs to watch out for and the things that could be dangerous for him/her?”

The ‘Pedestrian Challenge’ is an event organised by the city of Metz every year. It is open to all pupils in Grade 2 and Grade 3, i.e., several hundred children aged between 6 years-old and 8 years-old. The event consists of a life-size

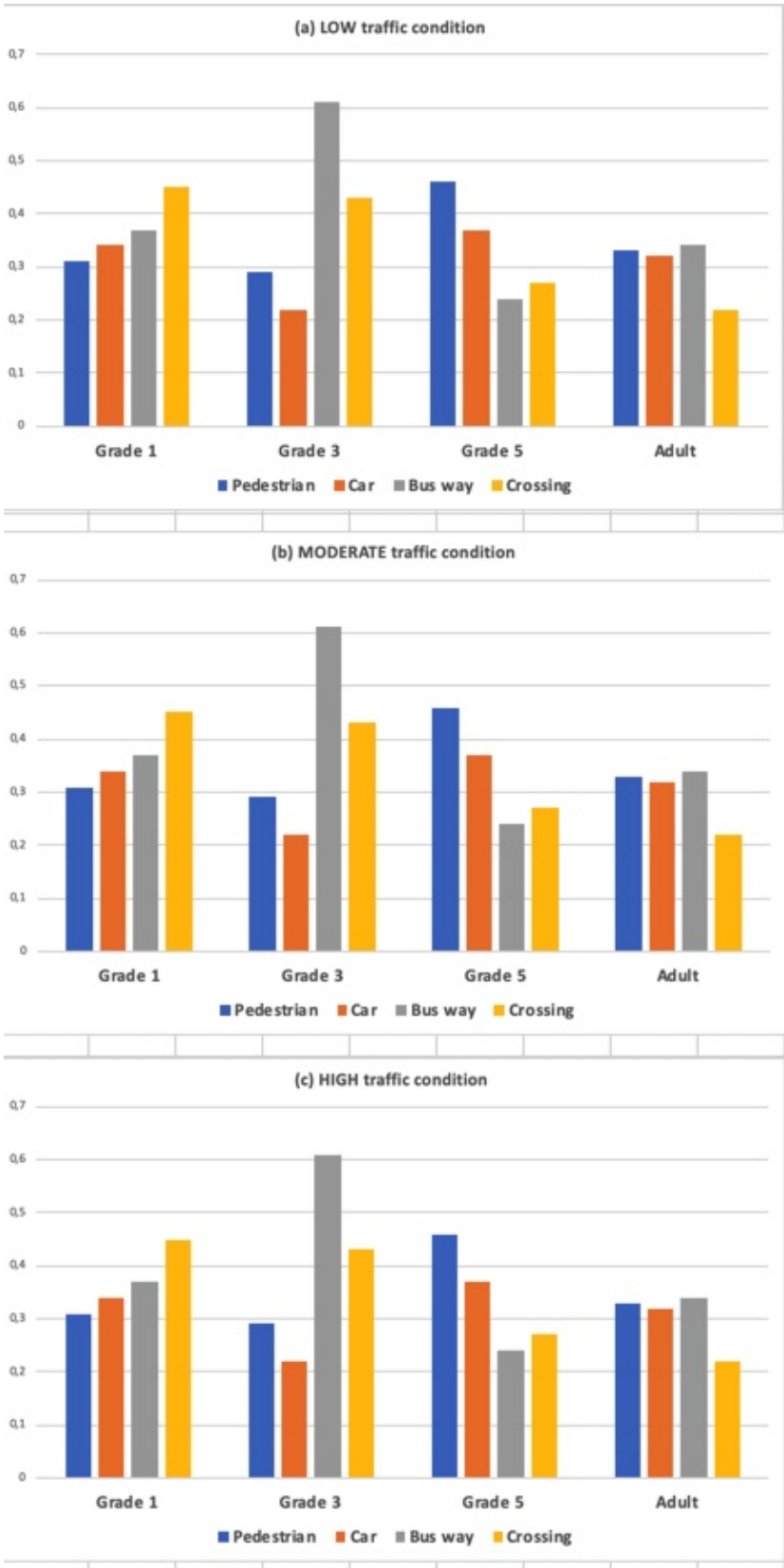


Figure 2. Mean of visual fixation duration for each Age Group (Grade 1, Grade 3, Grade 5, Adult), Traffic Condition (LOW / MODERATE / HIGH) for each Areas of Interests (AOI : Pedestrian / Car / Bus way / Crossing)

simulation of pedestrian navigation throughout the city. Pupils from Metz schools navigate independently along a predefined route through the city, confronting them with key pedestrian navigation situations (e.g., crossing several carriageways, bus lanes, cycle paths, complying with road safety rules, leaving car parks).

This individual task was carried out in the pupils' classrooms during their school time, in order to preserve a familiar context. Each child was given 15 minutes to draw his or her picture, during which time the experimenter collected verbalizations.

C. Analysis of drawings

In this study, we decided to take an objective quantitative approach to the qualitative material represented by these drawings. An ad hoc observation grid was therefore developed. The aim was to make a rigorous inventory of all the elements making up the drawings. The approach consists of adding each new element to the observation grid; the elements are then sorted and grouped into categories. The categories in the observation grid are characterised by their discriminating power and their exclusivity (an element cannot belong to several categories at the same time). Some categories are created to list elements that cannot be precisely identified or are irrelevant (e.g., the "decorative" category includes all decorative elements).

The observation grid developed to analyse the drawings produced by the 125 children ultimately comprised 7 categories:

- the point of view adopted by the child to produce the drawing (i.e., allocentric or egocentric orientation of the drawing [46]);
- decorative elements (trees, flowers, etc.);
- textual/language elements (e.g., instructions, first names, onomatopoeia);
- characters (male vs. female, and child vs. adult);
- traffic lanes (for car, pedestrian, bus, bike);
- signs (traffic lights for pedestrian, lights for bus, stop sign, crossing sign, pedestrian crossing);
- vehicles (car, bus, lorry, ambulance, fire engine, police vehicle, bicycle and other unidentifiable vehicles); For each of the items in each of the seven categories, the presence in each drawing produced was counted by distinguishing the objective accident risk in which each child is involved (low, moderate, high risk).

D. Main Results

With regard to the point of view adopted by the child in making the drawing (i.e., allocentric vs. egocentric drawing orientation), the egocentric perspective is in the majority. Figures 3 and 4 show two examples of drawings in these two categories. Between 60.53% and 90% of the drawings are produced from an egocentric point of view. The egocentric view is a 'first person' representation of the scene seen through the eyes of the author of the drawing. In other words, the elements represented in the egocentric drawing are drawn in the way they are actually perceived/seen by the children. In

contrast, the allocentric view, which concerns only a minority of the drawings, can be described as a 'bird's eye view'. In other words, it is a general view of the scene, as on a map, which enables the elements and the spatial links between them (e.g., distance, orientation, arrangement) to be identified.

Taking all categories of elements together, the average number of elements present in the drawings increases with the objective accident risk: drawings produced by children attending schools located in low-risk areas contain an average of 7.07 elements (SD = 2.90); those produced by children attending schools in moderate-risk areas have an average of 8.45 elements (SD = 3.17); finally, the drawings produced by children attending schools in high-risk areas have an average of 9.35 elements (SD = 3.17). The difference was significant only between low- and high-risk areas ($t(48) = 8.65, p = .001$). In any case, the richness of the drawings in terms of the number of elements present increases with the actual accident risk.

There were very few decorative elements in the drawings (e.g., trees, flowers, buildings). This relatively low proportion can certainly be explained by the time given to the children to produce their drawings (15 minutes). The textual and/or linguistic elements added to the drawings by the children were of two distinct types: some corresponded to material that children usually find in comics and children's literature (e.g., greetings, first names, dialogue in speech bubbles, onomatopoeia); the majority corresponded to instructions (e.g., "Look carefully to the left and right"). We note that none of the drawings produced by children attending schools in low-risk areas contain any textual elements. On the other hand, these textual/language additions were more frequent in the drawings produced by children attending schools in moderate (30.26%) and high (25%) risk areas.

The difference was significant only between low-risk areas and moderate- and high-risk areas (respectively: $t(95) = 19.26, p = .001$; $t(48) = 13.84, p = .001$). In other words, explicit instructions relating to urban navigation are more often present in the drawings produced by children attending school in areas of moderate or high accident risk.

As far as the characters are concerned, the drawings are very different depending on the schools where the children attend. In fact, only 13.79% of the drawings produced in low-risk areas include characters. Conversely, characters become a recurring feature in moderate-risk conditions (39.47%), and are very often present in high-risk conditions (65%). In the latter two cases, the drawings mainly represent scenes, most often a road crossing. The character is often of the same gender as the child producing the drawing (mechanism of identification and projection into the scene). The child then depicts himself alone or accompanied by her/his peers. The adults depicted are most often an instructor (a teacher?) or at least an adult performing traffic control and/or supervisory functions (presence of the "high visibility" jersey worn by security guards on the public highway). We can see that, although the instructions initially given to the pupils were intended to draw their attention to the vehicles, their drawings

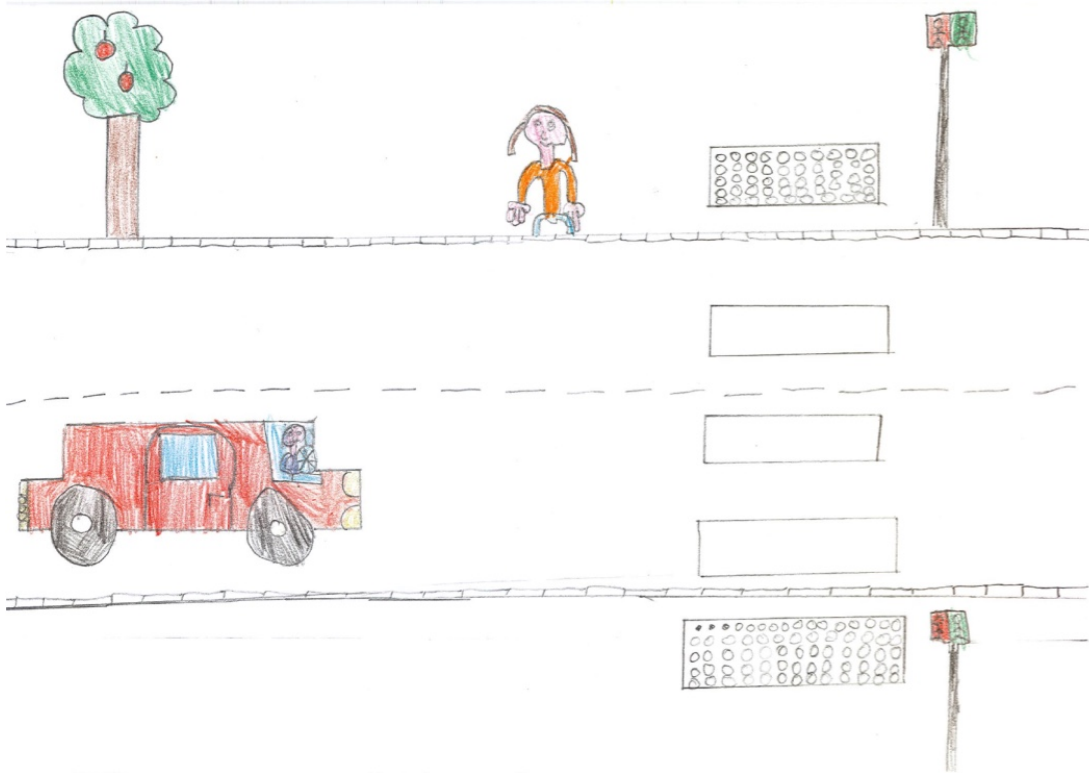


Figure 3. A first example of a drawing performed by one of our participant



Figure 4. A second example of a drawing performed by one of our participant

often include more characters than vehicles.

In the four traffic lanes identified in the drawings produced (pavement for pedestrians, bus/tram lane, cycle track and car lane), the 'car' lane is the most represented (from 27.59% to 65% of the drawings), followed by pavements (from 20.69% to 56.58%) and finally bus/tram lanes, which are the least represented (10.34% to 25%). It should be noted that cycle lanes are very rarely represented in the drawings (from 5.26% to 10%), whereas they are very present in the urban space that is so familiar to the participants in our study. Here again, the drawings differ greatly according to the schools in which the children attend: the number of traffic lanes represented increases for children attending schools located in moderate and high risk areas. In other words, even if this is only a statistically insignificant trend, the drawings produced by children attending schools in high-risk accident areas are becoming more complex and show multiple, multimodal roads (i.e., the presence of a bus/tram lane, a car lane and a cycle path on the same drawing).

Because road signs are such an integral part of our urban landscapes, and form the basis for learning the Highway Code and many of the prevention/awareness programs run in schools, it's only logical that children should frequently refer to road signs in their drawings. The main sign represented is the pedestrian crossing (up to 85 of the drawings include one). The pedestrian crossing therefore appears to be a central element, closely linked to the task of crossing the carriageway, which is of paramount importance to pedestrians. It should be noted that a large number of drawings are "collections" of signs in the sense that the traffic light (79.3%), pedestrian light (68.97%) and stop sign (65.52%) are all present. As with the road category, there is little signage for traffic lanes (bus, car, cyclist). On the other hand, there is no difference between the drawings produced by the children depending on the area in which they attend school, as far as these signs are concerned.

The category of vehicles is central to the study of representations of pedestrians, since they represent the main real danger to pedestrians. Although several types of vehicle were drawn by the children, the car was by far the most represented (from 45% to 65.79%). Buses and/or public transport are represented in more than a third of the drawings, a fact that is all the more interesting if we compare it with the low representation of bus lanes and associated signage. Generally speaking, the urban landscape in the drawings shows little variety of vehicles other than cars. Here again, there is no difference between the drawings produced by the children according to the area in which they attend school with regard to this signage.

E. Discussion

In this second study, we looked at the link between real (i.e., objectively established) accident risk and the representation of the urban environment in a population of child pedestrians aged 7 to 8, using drawings. Analysis of the 125 drawings collected yielded a number of interesting results.

The egocentric perspective is the perspective overwhelmingly present in the drawings produced by the children.

This result is consistent with the Piagetian approach to the development of intelligence, according to which children aged 7-8 (i.e., the period between the pre-operational stage and the concrete operations stage) understand the world from their own point of view. The drawings produced generally concern a specific road scene, namely crossing the carriageway. This is objectively the most dangerous activity and the one that children are most made aware of from an early age. From the point of view of the development of spatial representations, these evolve gradually, from a collection of elements present in the physical environment to the linking together of these elements [47], finally leading to a general and coherent spatial representation.

The children's drawings are more elaborate and richer (in terms of the number of elements represented) as the real accident risk in their environment increases. Not only does the number of elements in the drawings increase quantitatively, but these elements are more closely linked to each other, to the point where they form a real scene. The representation becomes more precise and the distinction between traffic lanes is more frequently observed in the drawings produced by children attending schools located in high-risk accident areas. The only exception is the signage, which is richer and more frequent in the drawings produced by children attending schools in low accident risk areas. We can assume that this latter representation corresponds to a representation of road safety rules (institutional knowledge).

In other words, the actual level of exposure to the risk of collision with motorised vehicles has an impact on children's mental representations in the sense that mental representations appear to be denser and richer in the most exposed children. Their representation then becomes a global and unified whole, which is reminiscent of the second stage of Endsley's model [48] According to Endsley, the first stage in risk perception is the perception of the constituent elements of the environment, giving rise to drawings that are juxtapositions of elements present in the environment, with no real link (i.e., a "collection of elements" such as a car, signs, a pedestrian crossing). The whole thing is not staged but simply presented on paper. However, "scene" type representations, the frequency of which increases with the real accident risk, go beyond this level to present the links between elements (e.g., a car is represented on a lane intersected by a pedestrian crossing, people waiting at the side of the road).

The participants in the second study were asked to draw a picture of the things to look out for when moving around the city. The majority of the drawings included people other than the author her/himself, usually peers, instructors or onlookers. It is interesting to note that the drawings include as many cars as characters, despite the fact that the risks of collisions are objectively associated with motorised vehicles. Analysis of the drawings also shows that other people (especially adults) are an important source of information when travelling in an urban environment. This importance of others when travelling in physical environments has already been noted by several authors: the behaviour of others informs us in which direction

to focus our attention [49] and enables us to anticipate certain behaviours of other pedestrians or drivers with whom the child shares the space [50]. In other words, as Granie points out [51][52][53], social norms condition us to share urban space by teaching us to interact with other users of this common space.

The development of risk representations should therefore be seen in the context of person-environment interaction: children are actors in their environment, acting on it just as the environment acts on them [54]. As [4] said, child pedestrian injury prevention can be characterized as a spectrum from active programs, designed to educate or train individuals to change their behavior, to passive interventions that increase the safety of products or environments in a manner that impacts all users. These "passive" interventions tend to be more costly and difficult to implement but are also more likely to result in real and sustained reductions in injury incidence.

Our results are in line with those which tend to show that it is not so much biological gender that determines risk perception as other factors such as the type of parental supervision according to gender or adherence to the gender stereotype [50].

While our study of mental representations (in this case, by means of drawings) has made it possible to identify certain relevant elements, it does not provide any information about the children's actual behaviour. So, as a way forward, we are already planning to study the link between the mental representations of child pedestrians and their actual behaviour when travelling in urban environments. Indeed, the link between 'perceived risk' and 'actual travel behaviour' is not all that logical: for example, a recent study [55] tends to show that, although a potential danger in their physical environment is correctly perceived and identified by young children (aged 7-8), they may adopt 'endangerment' behaviour (i.e., some children knowingly go into areas they perceive as dangerous). A combination of behavioural data and more subjective data is therefore desirable.

From a methodological point of view, although several data collection techniques can be used to investigate the mental representations of pedestrians (e.g., verbalizations concomitant with the activity of moving, verbalizations based on observation of static scenes), our study confirms that the use of drawing appears to be particularly suitable for questioning young people: it falls within their field of motor skills and particularly stimulates their interest. What's more, it's a technique that allows us to go beyond the limits of language, and is inevitably better suited to conveying a largely visual experience.

Finally, the results issued from our qualitative research highlights the interaction between children and their physical environment. The representations of subjects evolving in an environment richer in stimuli appear to be richer, denser and more interrelated. The elements represented are more numerous and more varied and, more importantly, the links between them are more frequent and better structured. Memory knowledge has a much more complex level of structuring, and this representation is more representative of individual

experience than institutional knowledge. In Endsley's model [48], this type of representation constitutes the intermediate stage of situational awareness in dynamic environments. In memory, this representation forms the basis of the judgments and predictions that enable the decision-making process. This is why the approach based on risk representation is inseparable from the study of pedestrian behaviour. Finally, the parental challenge is paradoxical: where is the happy medium between guaranteeing the child's well-being and developing autonomy?

IV. GENERAL DISCUSSION

From a theoretical point of view, our results obtained in our two studies show how the pedestrian's skills would be dependent on the development of at least two simultaneous capabilities: visual exploration strategy and cognitive processing abilities. First, the visual sampling strategies tend to be systematic in younger, not focusing on specific areas or strategic areas and, with age, the visual exploration strategy is specified and is interested in the peripheral areas of the visual field [33][45]. This development led to a more accurate and relevant information extraction from visual environment in urban areas [27][28][29]. Second, cognitive development allows greater information processing capacity [17][18][19], thus taking a more rapid and effective decision. From a theoretical point of view the use of poor visual strategy combined with a cognitive inability to process so many information that explains more time decision-making among young pedestrians in a dense traffic environment.

Several methodological limits prevent us to generalize the results obtained. First, the two experiments were conducted inside the school, which resulted in to cause a feeling of observation. The pupils often sought to provide "the good answer" whereas we are interested in their own answer. If the experiment were led in the school, it was a question above all of preserving a medium familiar and reassuring for the pupils. Second, stimuli used in our experiment were only visual and the information in peripheral vision necessarily decreased by the size of the screen. But, for ethical and technical reasons, it was not possible to carry out the experiment in real outdoor environment. Third, stimuli used in our first experiment were static (i.e., pictures): So, in our actual new studies, dynamic stimuli (i.e., videos) will be used to introduce dynamic factors, such as motion of vehicles and motion of other pedestrians in the scene. Moreover, even if visual information are crucial, we will add sounds in the experimental material to place participants in a more naturalistic setting. Our two studies tend to demonstrate on the one hand, that the development of pedestrian skills is essentially based on visual exploration of surrounding environment and on the other hand, these skills increase with the development of more general cognitive abilities, these two skills being crucial in the mental representation of children about hazards in their physical environments.

Nowadays, various techniques can be used to teach these skills, including classroom curricula, video or web-based instruction, real life practice with adult supervision and instruc-

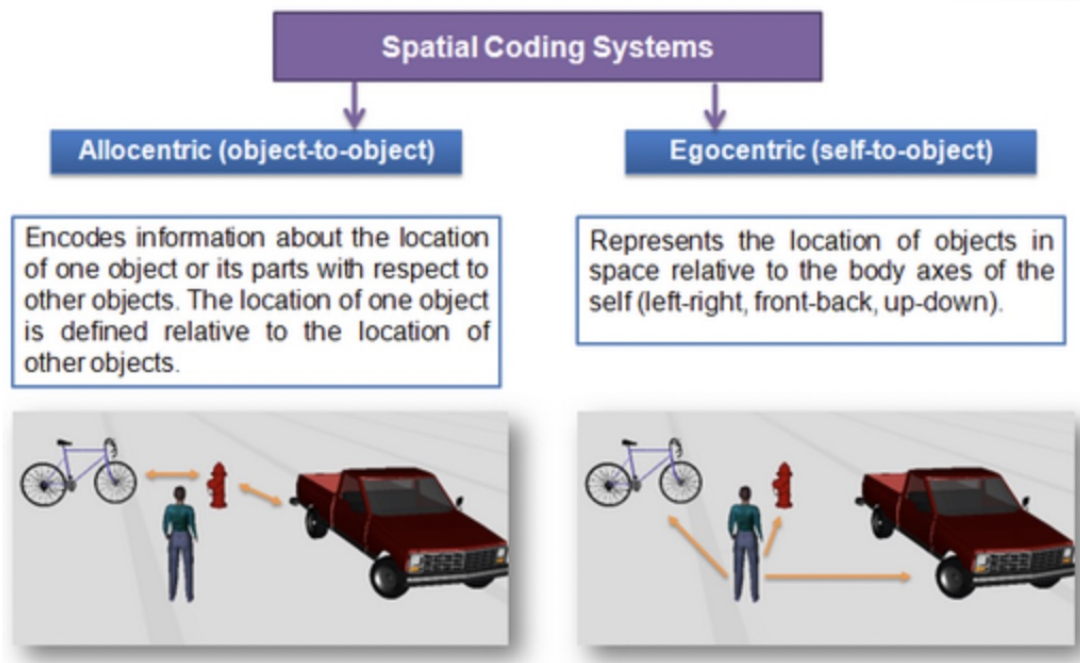


Figure 5. Allocentric vs. Egocentric Spatial Processing according to [45]

tion, and internet or smartphone-based virtual reality (VR) training [8][56][57][58].

V. CONCLUSION AND FUTURE WORKS

By using an experimental approach and eye-tracking techniques, our first study aimed to investigate the impact of one individual factor (Age) and one environmental factor (Traffic density) on three behavioural indicators related to competencies of very young pedestrians (aged 3-10 years): (i) the decision (i.e., "to cross" versus "not to cross the street"), (ii) the time spent in milliseconds to make this decision and (iii) the visual exploration of urban scenes displayed on pictures. This study is the first one to our knowledge which investigates visual exploration of urban scenes for very young children (under 4 years old). Using eye-tracking technique is interesting for several reasons. Visual exploration is an irrepensible behavior. Specifically, for young children with limited language capacity, the use of eye-tracking allows comparison with older children and adults. As we reported previously, to our knowledge, no study has looked at the visual exploration of such young pedestrians (under 4 years). Young audiences are more difficult to approach ethically. Younger cannot be put in a situation in real conditions, accompanied due to their motor skills development.

Our second study, based on drawings performed by young children, provided important findings about what is crucial for these young pedestrian in their physical environment. In other words, the first study offered quantitative and objective data while the second one offered qualitative and subjective data.

Although changes in policy, planning, and the built environment can have an important potential to improve young

pedestrian safety, many of the most commonly promoted strategies to address pedestrian risk focus on individual-level interventions to improve the skills or behavior of child pedestrians, their adult caregivers, or the drivers with whom they interact. Crossing the street is a complex task that requires the ability to identify a safe place to cross, recognize the relative speed of a moving car, judge the distance of the automobile, and quickly make a decision about whether it is safe to enter the roadway. So, all individuals must learn how to be a pedestrian. Rather than thinking of pedestrian safety as a body of knowledge to absorb, it is most useful to characterize it as a series of cognitive, perceptual, and motor skills that must be mastered to navigate the road traffic environment.

ACKNOWLEDGMENT

Thanks are due to all children, their parents and educational staff.

REFERENCES

- [1] S. Jordan et al., "Development of children's crossing skills in urban area: Impact of age and traffic density on visual exploration," in *COGNITIVE'2024*, 2024.
- [2] D. N. Lee, D. S. Young, and C. M. McLaughlin, "A roadside simulation of road crossing for children," *Ergonomics*, vol. 27, no. 12, 1984, pp. 1271–1281.
- [3] G. Lee, Y. Park, J. Kim, and G.-H. Cho, "Association between intersection characteristics and perceived crash risk among school-aged children," *Accident Analysis & Prevention*, vol. 97, 2016, pp. 111–121.
- [4] S. Kendi and B. D. Johnston, "Epidemiology and prevention of child pedestrian injury," *Pediatrics*, vol. 152, no. 1, 2023, p. e2023062508.
- [5] M. R. Zonfrillo, M. L. Ramsay, J. E. Fennell, and A. Andreassen, "Unintentional non-traffic injury and fatal events: Threats to children in and around vehicles," *Traffic Injury Prevention*, vol. 19, no. 2, 2018, pp. 184–188.

- [6] E. K. Adanu, R. Dzinyela, and W. Agyemang, "A comprehensive study of child pedestrian crash outcomes in Ghana," *Accident Analysis & Prevention*, vol. 189, 2023, p. 107146.
- [7] K. Koekemoer, M. Van Gesselien, A. Van Niekerk, R. Govender, and A. B. Van As, "Child pedestrian safety knowledge, behaviour and road injury in Cape Town, South Africa," *Accident Analysis & Prevention*, vol. 99, 2017, pp. 202–209.
- [8] J. A. Thomson, "Promoting pedestrian skill development in young children: Implementation of a national community-centered behavioral training scheme," *The Wiley handbook of developmental psychology in practice: Implementation and impact*, 2016, pp. 311–340.
- [9] M. Struik et al., "Pedestrian accident project report no. 4: Literature review of factors contributing to pedestrian accidents," Tech. Rep., 1988.
- [10] J. N. Ivan, P. E. Garder, and S. S. Zajac, *Finding strategies to improve pedestrian safety in rural areas*. United States. Dept. of Transportation, 2001.
- [11] C. DiMaggio and M. Durkin, "Child pedestrian injury in an urban setting descriptive epidemiology," *Academic Emergency Medicine*, vol. 9, no. 1, 2002, pp. 54–62.
- [12] W. W. Hunter, J. C. Stutts, W. E. Pein, C. L. Cox et al., "Pedestrian and bicycle crash types of the early 1990's," Turner-Fairbank Highway Research Center, Tech. Rep., 1996.
- [13] J. D. Demetre, "Applying developmental psychology to children's road safety: Problems and prospects," *Journal of Applied Developmental Psychology*, vol. 18, no. 2, 1997, pp. 263–270.
- [14] J. D. Demetre et al., "Errors in young children's decisions about traffic gaps: Experiments with roadside simulations," *British Journal of Psychology*, vol. 83, no. 2, 1992, pp. 189–202.
- [15] R. A. Schieber and N. Thompson, "Developmental risk factors for childhood pedestrian injuries," *Injury Prevention*, vol. 2, no. 3, 1996, p. 228.
- [16] J. L. Schofer et al., "Child pedestrian injury taxonomy based on visibility and action," *Accident Analysis & Prevention*, vol. 27, no. 3, 1995, pp. 317–333.
- [17] T. Pitcairn and T. Edlmann, "Individual differences in road crossing ability in young children and adults," *British Journal of Psychology*, vol. 91, no. 3, 2000, pp. 391–410.
- [18] A. Meir and T. Oron-Gilad, "Understanding complex traffic road scenes: The case of child-pedestrians' hazard perception," *Journal of Safety Research*, vol. 72, 2020, pp. 111–126.
- [19] A. Meir, T. Oron-Gilad, and Y. Parmet, "Are child-pedestrians able to identify hazardous traffic situations? measuring their abilities in a virtual reality environment," *Safety Science*, vol. 80, 2015, pp. 33–40.
- [20] J. M. Plumert, "Relations between children's overestimation of their physical abilities and accident proneness," *Developmental Psychology*, vol. 31, no. 5, 1995, p. 866.
- [21] J. M. Plumert and J. K. Kearney, "How do children perceive and act on dynamic affordances in crossing traffic-filled roads?" *Child Development Perspectives*, vol. 8, no. 4, 2014, pp. 207–212.
- [22] J. M. Plumert, J. K. Kearney, and J. F. Cremer, "Children's perception of gap affordances: bicycling across traffic-filled intersections in an immersive virtual environment," *Child Development*, vol. 75, no. 4, 2004, pp. 1243–1253.
- [23] D. C. Schwebel, A. L. Davis, and E. E. O'Neal, "Child pedestrian injury: A review of behavioral risks and preventive strategies," *American Journal of Lifestyle Medicine*, vol. 6, no. 4, 2012, pp. 292–302.
- [24] R. J. Brison, K. Wicklund, and B. Mueller, "Fatal pedestrian injuries to young children: A different pattern of injury," *Journal of Safety Research*, vol. 20, no. 1, 1989, pp. 41–41.
- [25] D. A. Sleet, M. F. Ballesteros, and N. N. Borse, "A review of unintentional injuries in adolescents," *Annual Review of Public Health*, vol. 31, 2010, pp. 195–212.
- [26] J. Warsh, L. Rothman, M. Slater, C. Steverango, and A. Howard, "Are school zones effective? an examination of motor vehicle versus child pedestrian crashes near schools," *Injury Prevention*, vol. 15, no. 4, 2009, p. 226.
- [27] M. Congiu et al., "Child pedestrians: Factors associated with ability to cross roads safely and development of a training package," Victoria: Monash University Accident Research Centre (MUARC), 2008.
- [28] J. A. Oxley, E. Ihssen, B. N. Fildes, J. L. Charlton, and R. H. Day, "Crossing roads safely: an experimental study of age differences in gap selection by pedestrians," *Accident Analysis & Prevention*, vol. 37, no. 5, 2005, pp. 962–971.
- [29] J. Oxley, B. Fildes, E. Ihssen, J. Charlton, and R. Day, "Simulation of the road crossing task for older and younger adult pedestrians: a validation study," in *ROAD SAFETY RESEARCH AND ENFORCEMENT CONFERENCE*, 1997, HOBART, TASMANIA, AUSTRALIA, 1997.
- [30] G. Simpson, L. Johnston, and M. Richardson, "An investigation of road crossing in a virtual environment," *Accident Analysis & Prevention*, vol. 35, no. 5, 2003, pp. 787–796.
- [31] M. L. Connelly, H. M. Conaglen, B. S. Parsonson, and R. B. Isler, "Child pedestrians' crossing gap thresholds," *Accident Analysis & Prevention*, vol. 30, no. 4, 1998, pp. 443–453.
- [32] M. L. Connelly, R. Isler, and B. S. Parsonson, "Child pedestrians' judgments of safe crossing gaps at three different vehicle approach speeds: A preliminary study," *Education and Treatment of Children*, 1996, pp. 19–29.
- [33] H. Tapiro, A. Meir, Y. Parmet, and T. Oron-Gilad, "Visual search strategies of child-pedestrians in road crossing tasks," *Proceedings of the Human Factors and Ergonomics Society Europe*, 2014, pp. 119–130.
- [34] S. Fotios, J. Uttley, and B. Yang, "Using eye-tracking to identify pedestrians' critical visual tasks. part 2. fixation on pedestrians," *Lighting Research & Technology*, vol. 47, no. 2, 2015, pp. 149–160.
- [35] T. Foulsham, E. Walker, and A. Kingstone, "The where, what and when of gaze allocation in the lab and the natural environment," *Vision Research*, vol. 51, no. 17, 2011, pp. 1920–1931.
- [36] A. C. Gallup et al., "Visual attention and the acquisition of information in human crowds," *Proceedings of the National Academy of Sciences*, vol. 109, no. 19, 2012, pp. 7245–7250.
- [37] D. Whitebread and K. Neilson, "The contribution of visual search strategies to the development of pedestrian skills by 4-11 year-old children," *British Journal of Educational Psychology*, vol. 70, no. 4, 2000, pp. 539–557.
- [38] R. Case, "The structure and process of intellectual development," *International Journal of Psychology*, vol. 22, no. 5-6, 1987, pp. 571–607.
- [39] R. Case, "Theories of learning and theories of development," *Educational Psychologist*, vol. 28, no. 3, 1993, pp. 219–233.
- [40] Z. Tabibi, K. Pfeffer, and J. T. Sharif, "The influence of demographic factors, processing speed and short-term memory on Iranian children's pedestrian skills," *Accident Analysis & Prevention*, vol. 47, 2012, pp. 87–93.
- [41] M. H. Matthews, *Making sense of place: Children's understanding of large-scale environments*. Barnes & Noble Books, 1992.
- [42] M. Farokhi and M. Hashemi, "The analysis of children's drawings: social, emotional, physical, and psychological aspects," *Procedia-Social and Behavioral Sciences*, vol. 30, 2011, pp. 2219–2224.
- [43] M. H. Matthews, "Young children's representations of the environment: a comparison of techniques," *Journal of Environmental Psychology*, vol. 5, no. 3, 1985, pp. 261–278.
- [44] R. T. Foley, R. L. Whitwell, and M. A. Goodale, "The two-visual-systems hypothesis and the perspectival features of visual experience," *Consciousness and Cognition*, vol. 35, 2015, pp. 225–233.
- [45] M. Kozhevnikov, S. Kosslyn, and J. Shephard, "Spatial versus object visualizers: A new characterization of visual cognitive style," *Memory & Cognition*, vol. 33, no. 4, 2005, pp. 710–726.
- [46] C. Martolini, G. Cappagli, E. Saligari, M. Gori, and S. Signorini, "Allocentric spatial perception through vision and touch in sighted and blind children," *Journal of Experimental Child Psychology*, vol. 210, 2021, p. 105195.
- [47] A. W. Siegel and S. H. White, "The development of spatial representations of large-scale environments," *Advances in Child Development and Behavior*, vol. 10, 1975, pp. 9–55.
- [48] M. R. Endsley, D. J. Garland et al., "Theoretical underpinnings of situation awareness: A critical review," *Situation Awareness Analysis and Measurement*, vol. 1, no. 1, 2000, pp. 3–21.
- [49] J. Kwak, H.-H. Jo, T. Luttinen, and I. Kosonen, "Modeling pedestrian switching behavior for attractions," *Transportation Research Procedia*, vol. 2, 2014, pp. 612–617.
- [50] B. K. Barton, "Integrating selective attention into developmental pedestrian safety research," *Canadian Psychology/Psychologie Canadienne*, vol. 47, no. 3, 2006, p. 203.
- [51] M.-A. Granié et al., "Qualitative analysis of pedestrians' perception of the urban environment when crossing streets," *Advances in Transportation Studies*, no. XXXI, 2013, pp. 14–17.

- [52] M.-A. Granié, F. Varet, and B. Degraeve, "The role of context and perception of road rules in the pedestrian crossing risky decisions: a challenge for the autonomous vehicle," in French-Japanese seminar on Simulation of On-Ground Mobility in Critical Situations: Cognitive Models and Computerized Modeling, 2019, pp. 221–233.
- [53] M.-A. Granié, M. Pannetier, and L. Gueho, "Developing a self-reporting method to measure pedestrian behaviors at all ages," *Accident Analysis & Prevention*, vol. 50, 2013, pp. 830–839.
- [54] M.-S. Cloutier, U. Lachapelle, and A. Howard, "Are more interactions at intersections related to more collisions for pedestrians? an empirical example in quebec, canada," *Transport Findings*, 2019.
- [55] J. Dinet, "Effect of hemophilia on risk perception in the outdoor activity of schoolers," *Enfance*, vol. 2, no. 2, 2015, pp. 199–223.
- [56] S. E. Bovis, T. Harden, and G. Hotz, "Pilot study: a pediatric pedestrian safety curriculum for preschool children," *Journal of Trauma Nursing—JTN*, vol. 23, no. 5, 2016, pp. 247–256.
- [57] D. C. Schwebel, L. A. McClure, and J. Severson, "Usability and feasibility of an internet-based virtual pedestrian environment to teach children to cross streets safely," *Virtual Reality*, vol. 18, 2014, pp. 5–11.
- [58] D. C. Schwebel et al., "Featured article: Evaluating smartphone-based virtual reality to improve chinese schoolchildren's pedestrian safety: A nonrandomized trial," *Journal of Pediatric Psychology*, vol. 43, no. 5, 2018, pp. 473–484.

Caption Generation for Clothing Image Pair Comparison Using Attribute Prediction and Prompt-based Visual Language Model

Soichiro Yokoyama
Faculty of Information
Science and Technology
Hokkaido University
Sapporo, Japan
email: yokoyama@ist.hokudai.ac.jp

Kohei Abe
Graduate School of Information
Science and Technology
Hokkaido University
Sapporo, Japan
email: ko.abe@ist.hokudai.ac.jp

Tomohisa Yamashita
Faculty of Information
Science and Technology
Hokkaido University,
Sapporo, Japan
email: tomohisa@ist.hokudai.ac.jp

Hidenori Kawamura
Faculty of Information
Science and Technology
Hokkaido University
Sapporo, Japan
email: kawamura@ist.hokudai.ac.jp

Abstract—Detailed information for product comparisons is necessary for consumers' purchasing process, especially during the information search and choice evaluation phases. However, conventional product descriptions, which are the primary source of information, tend to focus only on the product in question and thus do not adequately express the differences between products. Garments are treated as target products, and the content required to compare items is assessed from clothing comparison articles in lifestyle magazines. Two generation methods are proposed for comparison of a pair of garment items. The first method separately generates captions for each item and selects a caption pair that expresses differences. The other utilizes a Visual Language Model with a prompt designed based on the assessment. Subject experiments confirmed that the proposed Visual Language Model method accurately represented the feature differences between garments and provided helpful information for consumers to compare garments.

Keywords-consumer support; information provision; clothing caption generation; clothing attribute estimation, visual language model.

I. INTRODUCTION

This paper is based on the study presented initially at INTELLI 2024, The Thirteenth International Conference on Intelligent Systems and Applications [1]. An assessment of lifestyle magazine articles for clothing item comparison was added to organize the contents required in the captions. To generate captions that satisfy the requirements, a new method utilizing a Visual Language Model (VLM) was proposed, and its effectiveness was evaluated by comparison with the algorithm presented at the conference.

In the field of consumer behavior, the sequence of processes involved in the purchase of a product is widely recognized as the purchase decision-making process [2]. This process comprises five stages: problem recognition, information search, alternative evaluation, purchase decisions, and post-purchase evaluation. In the problem recognition phase, consumers identify their needs and problems, and collect information to

satisfy them in the information search phase. In the evaluation of alternatives, the consumer compares and evaluates products based on the collected information, and selects and purchases a specific product in the purchase decision stage. In the post-purchase evaluation, the degree of satisfaction was determined based on the results of the product use. During the information search and evaluation of alternatives phase, consumers need detailed information to understand the characteristics and differences of products and make the right choices. This information can originate from a variety of sources, such as user reviews, expert opinions, and comparison websites; however, product descriptions are one of the most important sources of information that consumers interact with in the early stages of their purchasing decisions. Product descriptions can successfully convey the basic features of a product; however, they tend to focus only on the product in question and do not adequately describe the differences between products. This lack of information may affect consumers' final purchasing decisions and post-purchase evaluations.

Image-caption generation is a research area for generating descriptive text from images; however, it primarily generates a single sentence for a single input image. It is impossible to generate a caption for each image by considering the relationships between multiple images. Some studies have aimed to generate distinctive image captions by comparing input images with similar images in a database; however, they cannot specify the images to be compared, as was the aim of this study. Recently, VLMs that receive prompt texts and images to generate text responses for general tasks have significantly improved and applied to the fashion domain. However, these models have yet to be utilized to generate such a caption pair.

This study aimed to provide adequate information to consumers when comparing products. As a concrete initial effort towards this goal, a method for generating captions that

highlight the differences between two products is proposed and evaluated. Clothing is selected as the target product. Clothing is an everyday purchase for consumers and has various features, such as pattern, material, length, and collar shape. Therefore, consumers need to compare product features during product selection. Articles from lifestyle magazine websites are gathered and analyzed to identify the key characteristics that the captions should include. Based on the assessment, this study considers captions for a pair of clothing items, generating one caption for each item that contains differences from the other item.

Two methods are proposed to achieve such a generation without requiring a large-scale dataset: caption pair selection and prompt-based VLM. The overview of each method is shown in Figure 1. In the caption pair selection method, two different garment images are independently input into an image-caption generation model to generate multiple captions. Next, the prominence of each attribute in each image is calculated using the garment attribute estimation model and the frequency of occurrence in the caption. This is compared between images, and the caption containing more salient attributes than one image is selected from the multiple captions generated for each image and output. The caption pair selection method yields captions that contain more salient features than one garment, with one sentence for each image and an average of approximately 14 words. Examples of the captions obtained are shown in Figure 1a. In the prompt-based VLM method, two garment images and carefully constructed prompts are given to a VLM, and the results are parsed to extract a caption pair. To fully cover the clothing attributes that should be included in the captions, chain-of-thought reasoning, where clothing attributes are first inferred, and captions are generated based on the attributes, is adopted.

In the subject experiment, it was evaluated whether the captions obtained using the proposed methods contained obvious errors, how well they described features that were only present in one garment, and whether they were useful for comparing garments. This experiment confirmed that the captions generated by prompt-based VLM adequately described the differences between products and provided useful information for product comparison.

The remainder of this paper is organized as follows. Section II describes work related to this study. Section III describes the proposed method. Section IV describes in detail the models and datasets used in the experiments. Section V describes the experiments on the comparative validation of the proposed method by employing different scoring methods. Section VI describes the experiments that qualitatively evaluate the captions generated by the proposed method. Finally, Section VII discusses the conclusion of this study and future perspectives.

II. RELATED WORK

This section describes the main areas relevant to this study, namely image caption generation, caption generation for multiple images, garment attribute estimation, and garment image caption generation.

TABLE I
COMPARISON OF IMAGE CAPTION GENERATION MODELS.

Model	BLEU4	METEOR
NIC [6]	27.7	23.7
NICA [9]	25.0	13.9
SCST [10]	31.9	25.5
ClipCap [11]	33.5	27.5
OFA [14]	44.9	32.5

A. Image Caption Generation

Image-caption generation is the task of generating an appropriate description of a single-input image. A comparison of the main image-caption generation models for the benchmark dataset Microsoft Common Objects in Context (MS COCO) [3] is presented in Table I. Bilingual Evaluation Understudy (BLEU) [4] and Metric for Evaluation of Translation with Explicit Ordering (METEOR) [5] are automatic metrics that measure the similarity between the generated and correct captions, with higher values indicating better model performance. Vinyals et al. [6] proposed a model based on a deep recurrent architecture that combines a Convolutional Neural Network (CNN) [7] and Long Short Term Memory (LSTM) [8]. Subsequently, Xu et al. [9] introduced an attention mechanism that focused on specific regions in an image when generating different words. Furthermore, Rennie et al. [10] proposed a model that incorporates reinforcement learning. Recently, image-language pre-training models that learn using large amounts of image-text pair data have achieved higher accuracy than conventional models. Mokady et al. [11] proposed a model that combines the image language pre-training model Contrastive Language-Image Pre-training (CLIP) [12] and the language model Generative Pre-trained Transformer 2 (GPT-2) [13], which reduces training time and achieves highly accurate caption generation. Wang et al. [14] also proposed a pre-training model using 20 million image-text pair data. All these models generate a single-sentence caption for a single input image. In this study, one-sentence captions are generated for each of the two input images. A one-input, one-output image caption generation model is used independently to generate multiple captions for each input image. Each caption is then scored, and the highest caption is generated one sentence at a time to generate a one-sentence caption for each of the two images.

Vision Language Models (VLMs) that receive arbitral text prompts and images and generate appropriate response text regarding the task specified in the prompt have recently achieved remarkable performance improvement. Inspired by the success of Large Language Models (LLMs) in conversation tasks realized with a large amount of training corpus and model parameters represented by GPT series [15], training generic VLMs that are capable of solving a variety of multi-modal tasks of vision and language have been developed. GPT-4 [16] was trained on image input in addition to text corpus, resulting in its capability in text generation through image recognition. Wang et al. [17] showed that a controllable

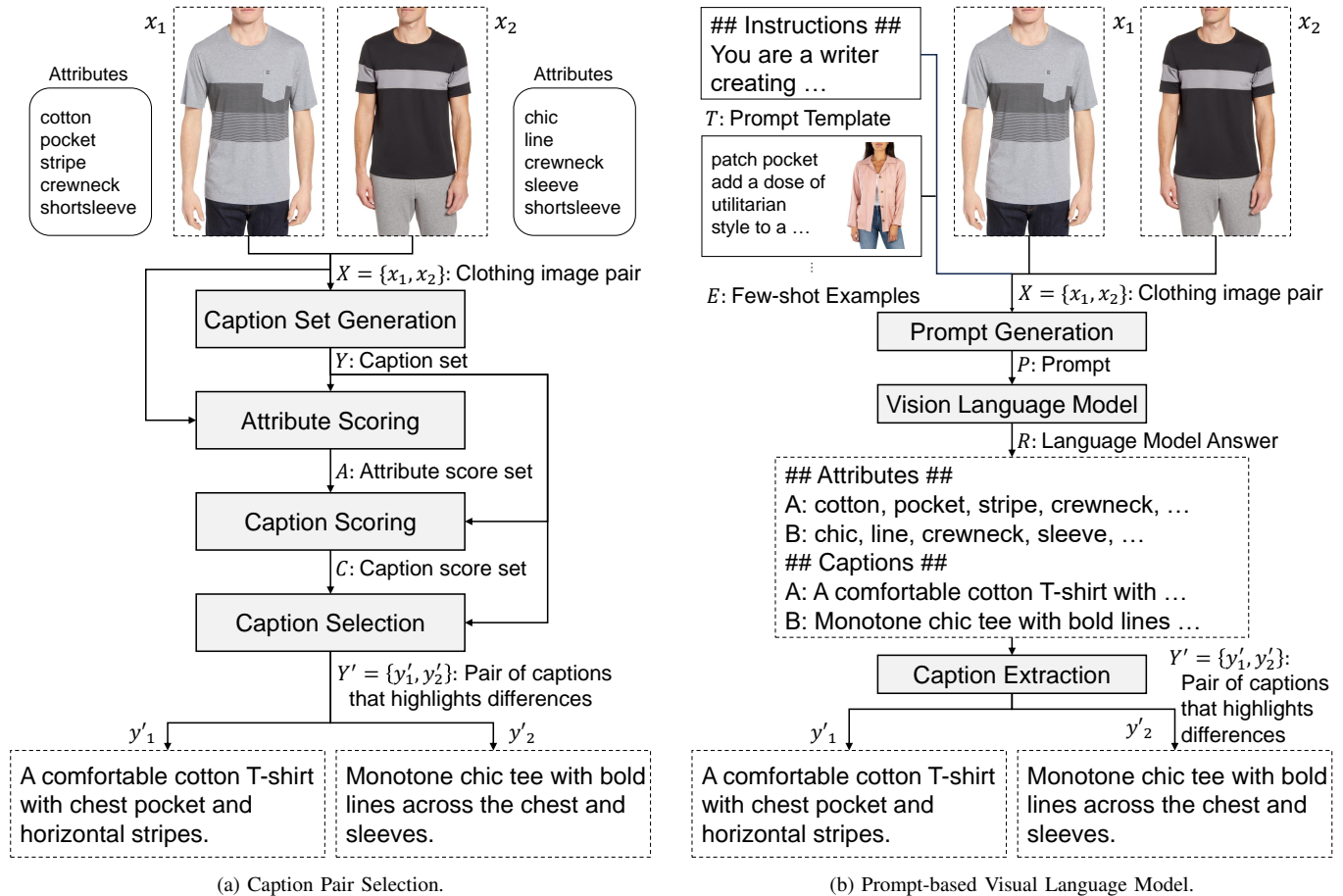


Figure 1. Overview of the proposed methods.

caption generation model can be obtained by training a VLM with a dataset containing multiple styles of captioning. In 2024, Ge et al. [18] presented a training-free pipeline that generates detailed captions by incorporating multiple VLMs and LLMs.

B. Caption Generation for Multiple Images

Several efforts have been made to generate captions for multiple images as an application of conventional image-caption generation. One example is the change in the image-caption generation initiative. This method identifies changes between two input images and generates a one-sentence caption describing the change [19][20]. In this study, a caption is generated for each input image. In conventional image-caption generation, which tends to generate generic sentences, the distinctive parts of the input images are often ignored. To address this problem, an approach called feature-based image-caption generation is currently in progress [21][22]. In this approach, a single input image is compared to a set of similar images in a database to identify the distinctive aspects of the input image, which are then reflected in the caption. However, this approach does not specify similar images explicitly. In this study, two specified images are compared. The attribute estimates calculated for each image are compared, and a

relative score is calculated. The caption score is then calculated by summing the attribute estimates that appear in the caption and is used for caption selection.

C. Clothing Attribute Estimation

Clothing attribute estimation is the task of estimating features, such as the material, pattern, collar shape, and sleeve length of clothing in an image. Examples of the estimated attributes include cotton, floral, sleeveless, and leather. This task has been applied to garment retrieval and recommendation. Chen et al. [23] proposed a model that combines a CNN [24] trained on a large image dataset, ImageNet [25] with a multilayer perceptron for a garment image retrieval task that matches images of garments worn by a person with those from a fashion e-commerce site. Similarly, Huang et al. [26] proposed a deep model that included two CNNs to handle street images and e-commerce site images in garment image retrieval. Both models were trained using bounding boxes to identify garment regions. In contrast, Liu et al. [27] proposed a model that learns garment landmark information, such as sleeve and collar positions, estimates the landmarks during inference, and uses this information as an aid for garment attribute estimation. A comparison of the garment-attribute estimation models on the benchmark dataset, DeepFashion

TABLE II
COMPARISON OF CLOTHING ATTRIBUTE ESTIMATION MODELS.

Model	Top-3 Recall	Top-5 Recall
WBIT [23]	27.46	35.37
DARN [26]	40.35	50.55
FashionNet [27]	45.52	54.61

[27] is presented in Table II. The Top-k Recall [28] was used as an evaluation metric. This assigns the top-k attributes with the highest probability of estimation to each image and measures the number of correctly estimated attributes. By estimating landmark information, FashionNet can better recognize the shape and position of garments and perform better than models that use only bounding boxes. Here, consumer perceptions of attributes are subjective and depend on age and gender. Different consumers may consider different attributes important when comparing garments. However, as a first attempt in this study, the weighting of the attributes did not change. Only estimates objectively calculated using the model were used.

D. Clothing Image Caption Generation

Sonoda et al. [29] proposed a method for searching for similar input images from a set of garment images they collected and applied the obtained garment information and features of similar images to a template. Yang et al. [30] proposed a framework that supports the creation of product introductions on e-commerce websites. In their study, attribute- and sentence-level rewards were introduced to improve the quality of captions generated. They also adopted a method for integrating the training of the model using maximum likelihood estimation, attribute embedding, and reinforcement learning. In addition, a large dataset for garment image-caption generation containing approximately one million images was constructed. Cai et al. [31] removed noisy garment images and reconstructed a clean garment image dataset. These studies generated captions describing the salient features of a single-input garment image.

VLMs and LLMs are utilized in the fashion domain to provide more user-friendly interfaces for practical tasks such as retrieval and report generation. Chen et al. [32] integrate ChatGPT with a fashion retrieval system for understanding user queries. Ding et al. [33] propose a system that produces reports by analyzing catwalk images on fashion shows using a VLM. Maronikolakis et al. [34] evaluated the effectiveness of publicly available LLMs as a conversational agent in the fashion domain and showed promising results with GPT-4, a state-of-the-art LLM.

They are insufficient for the purpose of this research, that is, to provide information when comparing garments, in that they cannot express the detailed differences between different garments. In this study, a caption is generated that highlights the differences between two input garment images.

III. PROPOSED METHOD

This section describes the clothing caption that highlights the differences between garment image pairs and the genera-

tion methods proposed in this study. First, real-world clothing captions that describe differences between garment pairs are assessed. Possible generation approaches are considered, and two promising methods implemented and evaluated in this paper are selected. The first method, caption pair selection, uses a conventional clothes attribute estimator and a clothes caption generator trained on an existing clothes dataset to sample caption candidates and select a pair that contains the most different attributes. The other utilizes a publicly available vision language model with a specifically designed prompt to generate differentiating caption pairs. Both proposed methods are trained on the existing clothing dataset or the general text corpus, requiring little to no additional dataset to generate differentiating caption pairs.

A. Generation of Clothing Captions that Highlight Differences

We collected articles from lifestyle magazine websites where multiple clothing items of the same category were compared for customers considering purchase and assessed clothing attributes discussed in the articles to determine the captions to be generated in the proposed method. An instance of clothing captions that highlight differences between a pair of clothing images is produced based on the articles. Finally, algorithms for generating the captions are discussed.

1) *Assessment of Magazine Articles on Garments Comparison*: Eight articles comparing multiple clothing items are collected from Japanese lifestyle magazine websites to assess their comparison approach and item attributes mentioned. Each article is published by lifestyle magazines and compares clothing items of the same category with similar price ranges, typically from different manufacturers, to provide customers with information about each item. Since the original magazine articles were written in Japanese, all the articles were translated into English for assessment. Table III summarizes translated titles, the number of items compared, the description approach for items, and discussed attributes in each article. The number of clothing items compared in one article ranges from 2 to 6, and four out of eight articles compare two items.

Approaches to describe differences among the clothing items can be divided into two categories. One comprises multiple paragraphs, each explaining the items' features on a specific attribute. The other consists of paragraphs explaining each item, enumerating notable attributes. For example, the former first states the design features of two items and their differences, such as the number and location of pockets. The difference of materials in a subsequent paragraph. In contrast, the latter describes the design and material features of one of the items. The design and coordination recommendation for the other item is discussed in successive paragraphs. The highlighting of differences is evident in the former as the differences are discussed per attributes contributing to understanding the breakdown of differences. In the latter, some articles contain sentences that explicitly state the difference from other items, such as "This one is smoother to touch." and "From the four tried-on items, this one was felt to fit my feminine style the best." for highlighting the differences.

TABLE III
ASSESSMENT RESULT OF CLOTHING COMPARISON ARTICLES FROM LIFESTYLE MAGAZINE WEBSITES.

No.	Title	Items	Category	Description approach	Discussed attributes
1	GU's "Tucked Wide Pants" for a beautiful look. Comparison with the one from UNIQLO	2	Pants	Per attribute comparison	Design, Silhouette, Size, Material, Comfort, People recommended for
2	[A Thorough Comparison of the Most Cost-effective Products] Which is superior, UNIQLO or MUJI? Comparison of "linen shirt" priced at 3,990 yen	2	Tees	Description of each item and per attribute comparisons	Design, Material
3	[Workman VS North Face] Which cardigan to choose for autumn/winter? An enthusiast explains the recommendation!	2	Sweater	Description of each item with notable differencing points	Design, Material, Effect, Coordination, People recommended for
4	[Comparison Report] What are the differences between UNIQLO and GU's much talked-about "parachute cargo pants"?	2	Pants	Description of each item and per attribute comparisons	Design, Silhouette, Material, Impression, Effect, Coordination, Wearing scene
5	[Workman, UNIQLO, Muji] A thorough comparison of the "best men's T-shirts"! Which is the recommendation available under 2,000 yen?	3	Tee	Description per item with differences and similarities	Silhouette, Size, Material, Color, Impression, Effect, Wearing scene, People recommended for
6	Thorough Comparison of UNIQLO's 2024! "White T-shirts" that will be very useful this summer are here!	4	Tee	Description per item with unique feature of each item	Design, Silhouette, Size, Material, Impression, Coordination, People recommended for
7	A hot topic on SNS! Comparing 4 pairs of UNIQLO cargo pants. Which one is the best fit for you?	4	Pants	Description per item with unique feature or differences	Design, Silhouette, Material, Color, Impression, Wearing scene, Coordination, People recommended for
8	For those who can't choose from too many "UNIQLO White T's". Comparison of 6 models including the most popular No.1 and men's [Try-on review]	6	Tee	Description per item with differences	Design, Silhouette, Material, Impression, Effect, Wearing scene, Coordination, People recommended for

TABLE IV
TEXT LENGTH OF MAGAZINE ARTICLES.

No.	Items	Average count for each item	
		Sentences	Words
3	2	4.5	78.0
5	3	5.3	110.3
6	4	4.0	79.5
7	4	5.3	90.8
8	6	5.8	110.8

This article considers caption generation for a pair of clothing images representing differences between items. A pair of captions are generated, each describing the corresponding item. This approach is commonly utilized in 5 out of 8 articles assessed. A comparison of only two items is considered to validate the basic feasibility and usefulness of caption generation that highlights differences. Note that all the articles assessed for more than three items employ this approach, which implies its extendability.

For the articles that explain each garment in different paragraphs, the text length for one item was approximately five sentences and 100 words in English. Table IV shows the average number of sentences and words. Each item was explained in about five sentences and 100 words for all the articles assessed. Therefore, the generated caption length for each item is also targeted at five sentences and 100 words.

We have organized clothing attributes commonly used in the articles. The following attributes were used to characterize garments, alone or in combination with others.

Design, Silhouette, and Details Shape of garments such as

V-neck and crewneck, clothing size outlines of the wearers like loose-fitting, tapered hem and smooth fit over the shoulders, and decoration or utility details such as the number and locations of pockets, ribbons, and straps.

Material Clothing materials and textures such as linen, hemp, glossy finish, and smooth texture.

Color, Pattern, and Print Available colors, patterns, and garment prints. e.g., red, solid color, bright color, and logos.

Additionally, the following derivative attributes were described in conjunction with the attributes above, often intended to provide solid evidence for more subjective derivative attributes with objective attributes.

Impression Impressions that others may receive from the wearer when wearing the garment. e.g., the material with a light sheen gives it a high and beautiful look.

Effect The effect of wearing the garment on the wearer's body shape and comfort. e.g., hip-hugging length for a slimming effect, soft against the skin with cotton blend material for a non-stress fit

Wearing Scene Situations where it is assumed that wearing clothing is appropriate and effective. e.g., the slightly longer sleeves are also perfect for the morning and evening temperature differences and the chilly rainy season.

Coordination and Styling Recommendation on other items that suit the garment. e.g., this shirt is not too long and can be easily matched with tapered silhouette pants

These derivative attributes were often written as recommendations for those with specific ideas, such as "The stretchy

cotton dobby material has a firm feel, making it ideal for those who want to enjoy an elegant look.” possibly because of the subjective nature of these attributes.

Reflecting on the assessment above, we produced an example of a pair of captions that describe the garment of two images as shown in Figure 2. Each caption contains the discussed attributes, explicitly comparing with the other garment and recommending specific people. The part of the text in the figure that refers to attributes is shown in bold. The part that compares with the other item is underlined. Each image of the example is obtained from the official website of UNIQRO and trimmed by the authors to align with other clothing images included in the dataset. The image for Garment A is taken from <https://www.uniqlo.com/jp/ja/products/E468503-000/00>, and <https://www.uniqlo.com/jp/ja/products/E472071-000/00> for Garment B. The caption refers to the other garment as “Garment A/B” for generality. Note that simple algorithms, such as substitution with product names, can easily alter this behavior.

2) *Possible Algorithms for Caption Generation*: We discuss possible algorithms for generating captions that highlight differences. Due to the lack of an existing dataset for such captions, this paper considers two algorithms: caption selection and a prompt-based VLM. First, existing caption generation models for garments are evaluated. Differentiation approaches for a pair of input garments are discussed. Finally, overviews of the two proposed methods are presented.

The caption selection method uses existing captioning models that generate captions from a single clothing image. Firstly, candidates of captions are separately generated for each clothing image of a pair using those models. Then, an additional algorithm selects the most appropriate pair. We propose a selection method based on an existing clothing attribute estimator. Another captioning model that accepts a clothing image and specific attributes to include in the caption could be used for greater efficiency. However, since the pre-trained weights of this model are not publicly available at the time of writing, such an approach is omitted from this paper.

VLMs are trained to generate response text based on text prompts and images for general tasks. With an appropriate prompt, these models can be used to generate captions that highlight differences. We propose a prompt-based VLM method for the caption generation. Most advanced VLMs can be accessed with API, allowing users to input custom text prompts and images and return a responding text. Preliminary experiments are conducted on three VLMs with publicly available APIs, and the most effective model is selected. With an abundant dataset of differentiating captions, VLMs could be fine-tuned for potentially more precise and context-aware caption generation.

B. Caption Pair Selection Method

An overview of the method is presented in Figure 1a. The method considers a pair $X = \{x_i \mid i = 1, 2\}$ of different garment images as input and outputs a caption pair $Y' = \{y'_i \mid i = 1, 2\}$ corresponding to each image, where x_i

is the i -th garment image, and y'_i is the output caption corresponding to x_i . In Figure 1a, the attribute set annotated to the image is displayed next to each image. This method comprises four modules: caption set generation, attribute scoring, caption scoring, and caption selection. The following sections describe these modules in detail.

1) *Caption Set Generation Module*: The caption set generation module considers a pair X of different garment images as input, inputs each image independently of the image caption-generation model, and outputs a caption set $Y = \{y_{ij} \mid i = 1, 2; j = 1, 2, \dots, J\}$ corresponding to each image. Here, y_{ij} represents the j -th caption for image x_i . The image-caption generation model used in this study is described in detail in Section IV.

2) *Attribute Scoring Module*: The attribute scoring module considers a pair of different garment images X and a caption set Y as input and outputs a set of attribute scores $A = \{a_{ik} \mid i = 1, 2; k \in K\}$ for each image. Here, K is the set of attributes to be evaluated and a_{ik} is the score of attribute k for image x_i . An attribute score is a numerical expression of the prominence of a particular attribute exhibited by a garment image; the higher the score, the stronger the garment image that exhibits that attribute. An example of an attribute score for the garment image x_1 in Figure 1a is 0.20 for crewneck, 0.15 for pocket, and 0.01 for sleeveless, which were calculated to be higher when the image had the attribute prominently and lower when it did not. In this study, two methods of attribute scoring were considered: attribute scoring based on attribute estimation, and attribute scoring based on frequency of occurrence.

Attribute scoring based on attribute estimation uses a garment-attribute estimation model, whose output is the estimated probability of each attribute for an input-garment image. The estimated probability of an attribute for each image was calculated, and this value was used as the attribute score. This is illustrated in (1), where p_{ik} is the estimated probability of attribute k for image x_i . The clothing attribute estimation model used in this study is described in detail in Section IV.

$$a_{ik} = p_{ik} \quad (1)$$

Attribute scoring based on frequency of occurrence assumes that the caption generated for each garment image using the caption set generation module reflects the garment characteristics. If a particular attribute appears frequently in a caption set, it can be regarded as one of the main features of the garment. This method calculates the frequency of occurrence of each attribute in the caption set for each image and uses this value as the attribute score. This is illustrated in (2), where f_{ijk} is the number of occurrences of attribute k in the caption y_{ij} .

$$a_{ik} = \frac{1}{J} \sum_{j=1}^J f_{ijk} \quad (2)$$

3) *Caption Scoring Module*: The caption scoring module considers a caption set Y and an attribute score set A as inputs, and outputs a caption score set $C = \{c_{ij} \mid i = 1, 2; j =$



This **dark brown, simple cotton T-shirt** features a **rounded neckline** that creates a more feminine look compared to Garment B. The **soft fabric** and **slightly slim silhouette** are distinctive, not only giving a **slimmer appearance** but also making it **perfect for creating a polished T-shirt style**, such as **layering it under a jacket**. With fewer decorations than Garment B, it has a more **refined appearance**, making it ideal for **daily outings and office casual wear**. Recommended for those who want to **project an air of elegance in a T-shirt outfit**.

Bold text indicates attributes and underlined text indicates comparison.

(a) Garment A and corresponding caption.



This **taupe T-shirt** features **ruffled sleeves, which Garment A lacks**, adding not only **elegance** but also effectively **covering the upper arms**. The fabric has a **smooth texture**, making it **comfortable to wear even in summer**. The **slim design prevents it from bunching when tucked in**, which is a great advantage. It looks **stylish** on its own and **pairs well with skirts**, making it perfect for **dates and daily outings**. With a more girly impression than Garment A, it's recommended for those who find **plain T-shirts too simple**.

Bold text indicates attributes and underlined text indicates comparison.

(b) Garment B and corresponding caption.

Figure 2. Examples of captions for a pair of clothing images. (Images are taken from the UNIQLO official website and trimmed by the authors)

$1, 2, \dots, J\}$. The caption score is a numerical expression of the extent that the caption reflects the salient attribute differences between the garment images and attributes specific to each image; a higher score is regarded as emphasizing the differences between one image and the other. Here, c_{ij} represents the score of the caption y_{ij} . In this study, two caption scoring methods were considered: caption scoring based on the comparison of top attributes and caption scoring based on the addition of relative scores. These methods are described in detail as follows.

Caption scoring based on comparison of top attributes first obtains an attribute set K_i^{top-n} with the top n attribute scores for each image. Next, the difference set D_i of K_i^{top-n} for each image is the difference attribute set, and the product set T is the common attribute set. These are presented in (3)~(5):

$$D_1 = K_1^{top-n} \setminus K_2^{top-n} \quad (3)$$

$$D_2 = K_2^{top-n} \setminus K_1^{top-n} \quad (4)$$

$$T = K_1^{top-n} \cap K_2^{top-n} \quad (5)$$

Finally, the difference between the number of attribute occurrences in the different attribute sets and the number of attribute occurrences in the common attribute set for each caption was calculated and used as a caption score. This process is illustrated in (6), where f_{ijk} is the number of occurrences of attribute k in caption y_{ij} .

$$c_{ij} = \sum_{k \in D_i} f_{ijk} - \sum_{k \in T} f_{ijk} \quad (6)$$

This method assigns higher scores to captions containing more differentiated and fewer common attributes.

Caption Scoring Based on Relative Score Addition first calculates the difference in attribute scores between images

to obtain the relative attribute scores Δa_{ik} . These are given by Equations (7) and (8), respectively.

$$\Delta a_{1k} = a_{1k} - a_{2k} \quad (7)$$

$$\Delta a_{2k} = a_{2k} - a_{1k} \quad (8)$$

Next, the relative attribute scores corresponding to the attributes in the caption are added and used as the caption score. The process is described in (9), where $K_{y_{ij}}$ represents the set of attributes contained in the caption y_{ij} .

$$c_{ij} = \sum_{k \in K_{y_{ij}}} \Delta a_{ik} \quad (9)$$

Using this method, captions containing more attributes with relatively high attribute scores have higher scores.

4) Caption Selection Module: The caption selection module considers the caption sets Y and C as input, selects the caption with the highest caption score in the caption set corresponding to each image, and outputs a set of captions $Y' = \{y'_i \mid i = 1, 2\}$ that highlights the differences. This process is represented by (10).

$$y'_i = \underset{y_{ij}}{\operatorname{argmax}} c_{ij} \quad (10)$$

C. Prompt-based Visual Language Model

The method utilizes a VLM, which receives text prompts and images to generate responding text for general tasks. To acquire appropriate captions highlighting differences between clothing images, prompting techniques of few-shot examples and chain-of-thought reasoning are adopted.

An overview of this method is shown in Figure 1b, formulated as follows. First, a pair $X = \{x_i \mid i = 1, 2\}$ of garment images is given. Combined with a text prompt template T and few-shot examples $E = \{x_i \mid i = 1, 2\}$, inputs for VLM

are formatted as P in the prompt generation module. With chain-of-thought reasoning, VLM first predicts the attributes of the garments to improve the coverage and then generates captions. Therefore, the VLM response R contains attributes and captions for both garments. The caption extraction module separates the desired caption pair $Y' = \{y'_i \mid i = 1, 2\}$ corresponding to each image from the rest of R .

1) *Prompt Generation Module*: Prompts to the VLM significantly impact the quality of the generated caption and thus are carefully constructed with chain-of-thought reasoning and few-shot examples. Prompt template T specifies an instruction for caption generation, constraints on the output of each step of chain-of-thought reasoning should follow, followed by few-shot examples E that instantiate the desired generation contents and formats for specific inputs, and finally, a format that the response should follow and the input images. By concatenating the contents, the prompt generation module produces prompts P .

The contents of the prompt template T are as follows. Since VLM APIs utilized in this paper have a separate system input text that controls the role of the VLM in addition to user input texts, the template consists of system and user text parts.

The system input text specifies instructions for the VLM. The writer's role in an e-commerce site that compares two garments is given. Generation is structured in two steps. First, each garment image's previously discussed clothing attributes are inferred as a text of attributes enumeration. Then, a pair of captions is generated based on the input images and the results of the first step.

The user input text includes specific output constraints for each step. For the first step, garment features are categorized, and explanations and examples for each category are given. Constraints for the second step state the text length and the attributes that should be included in the captions. The example of input image pairs and corresponding output texts follows. Then, an output format is specified in JSON for both steps for easy extraction. Finally, clothing images of the pair X are appended to the prompt.

The few-shot examples E are produced by the authors regarding the assessment of the magazine articles. Two pairs of clothing items with images, corresponding attributes, and captions are presented, which consist of a pair of shirts, shown in Figure 2, and a pair of pants.

2) *Visual Language Model*: The constructed prompt P is given to the VLM, and the response R is returned. In principle, any VLM that accepts arbitrary images and text prompts can be used. For caption generation, however, the VLM has to be flexible enough to follow the instruction of P and produce a valid JSON formatted response that contains two corresponding captions for each input.

3) *Caption Extraction Module*: The response R from the VLM is formatted as text data, which can be parsed as valid JSON. The response should contain the estimated attributes for each image as the first step, followed by a corresponding caption for each item as the result of the second step. By accessing the relevant elements of the interpreted JSON object

from R , the caption extraction module extracts $Y' = \{y'_i \mid i = 1, 2\}$ from R .

The extraction of captions is straightforward, provided that the response R adheres to the format specified in the prompt P . In our experiments, the VLM consistently produced responses in the expected format, thus eliminating the need for additional post-processing algorithms. Since most VLMs generate responses probabilistically in an auto-regressive manner, the generation process can be repeated if the initial response does not conform to a valid JSON format. This approach ensures that the final output meets the required structure.

IV. MODELS AND DATASETS

This section describes the image-caption generation models, garment attribute estimation models, garment image datasets, and VLMs used in the study.

A. Image Caption Model and Clothing Attribute Estimation Model

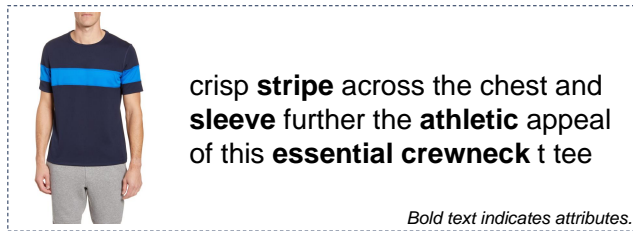
This study is looking at reflecting different national and regional fashion cultures in captions in the future. Therefore, image-caption generation models that can handle garment image data in various languages are desirable. Among the image-caption generation models compared in Section II, ClipCap [11] is a combination of CLIP and the language model GPT-2. It is easy to handle non-English data because CLIP exists for multiple languages [35], and the language model Generative Pre-trained Transformer 4 (GPT-4) [16], which is similar to GPT-2, supports multiple languages. Furthermore, as shown in Table I, the accuracy is sufficiently high among the major image-caption generation models. Therefore, in this study, ClipCap was used as the image-caption generation model in the caption set generation module. FashionNet [27] was used as the garment-attribute estimation model in the attribute-scoring module. This model estimates the landmarks of a garment and uses the obtained information for garment attribute estimation. This model can capture the fine-grained features of a garment image and is highly accurate.

B. Clothing Image Dataset

A comparison of the main garment image datasets is shown in Table V. In this study, the FACAD170K garment image dataset [31] with both attributes and captions, which enables an attribute-based caption evaluation, was used to train the image-caption generation model. An example of the FACAD170K data is shown in Figure 3. Each garment image was crawled from a generic website, mainly Google Chrome, and was either an image of a person wearing the garment or an image of the garment alone, with a one-sentence caption from the web. The data collected using this method reflect the variety of styles and trends in clothing that real consumers interact with on a daily basis and are therefore highly suitable for simulation and analysis to mimic the context of consumers' clothing choices. The same caption is provided for garments of different colors. The bold text in the captions for Figure 3 represents multiple attributes assigned to a single garment image. FACAD170K

TABLE V
COMPARISON OF CLOTHING IMAGE DATASETS.

Dataset	Number of images	Attributes	Captions
FACAD170K [31]	178,862	yes	yes
DeepFashion [27]	289,222	yes	no
FashionGen [36]	325,536	no	yes
iFashion [37]	1,062,550	yes	no



(a) Clothing image A and corresponding caption.



(b) Clothing image B and corresponding caption.

Figure 3. Examples of data from the FACAD170K dataset.

TABLE VI
HIGH-FREQUENCY ATTRIBUTES COMMON TO BOTH FACAD170K AND DEEPFASHION.

Attribute	Frequency (%)
cotton	4.53
cut	4.41
soft	3.76
sleeve	2.98
fit	2.81
leather	2.58
stretch	2.46
classic	2.45
knit	2.31
strap	2.25

has 990 attributes. In contrast, training the garment-attribute estimation model requires bounding boxes and landmark information to identify garment regions. However, FACAD170K did not contain these annotations. Because annotation is time-consuming, we used FashionNet's DeepFashion [27] pre-training model for garment attribute estimation. DeepFashion contains 1000 attributes, 292 of which match FACAD170K. The top ten attributes with the highest frequency of occurrence in FACAD170K and their frequencies are listed in Table VI. FACAD170K and DeepFashion data with these attributes were used to evaluate the proposed method.

C. Visual Language Model

In the prompt-based VLM method, a VLM must follow a complex introduction and produce a result in a valid JSON format. We compared three state-of-the-art models that are publicly available through APIs and satisfy these requirements. They achieved competitive results with the vision-text benchmark, and it was determined that there was a need to compare performance on the tasks in this study.

OpenAI GPT-4o [38] A flagship model from OpenAI with multi-modal ability when the experiments are performed.

Anthropic Claude 3.5 Sonnet [39] The latest model from Anthropic at the time of the study.

Google Gemini 1.5 Pro [40] A VLM model from the Google Gemini Team reported the best performance on vision benchmarks across their models during the study.

Although models with publicly available weights, such as LLaVA [41], are attractive options since fine-tuning and integration with other models are practical, these models tend to perform inferior to the compared models in the fashion domain in zero-shot or few-shot settings. Therefore, they are excluded from comparison in this study.

V. COMPARATIVE VERIFICATION OF MODULES IN THE PROPOSED METHODS

Each module in the proposed method is compared and validated in this experiment. Multiple algorithms or models are presented for some modules in the proposed methods. To identify the most suitable algorithm for each module, preliminary experiments were conducted before engaging in the more labor-intensive qualitative evaluation of the generated captions.

A. Caption Pair Selection Method

This experiment aimed to compare and validate attribute scoring based on attribute estimation and frequency of occurrence in the attribute scoring module and caption scoring based on the comparison of top attributes and the addition of relative scores in the caption scoring module to find the best combination of methods for generating captions that highlight differences.

1) *Methods*: In this experiment, the captions generated using the four proposed methods were automatically evaluated. In the caption set generation module, the image-caption generation model ClipCap was trained using 168,862 training data points from FACAD170K. The key parameters during training were set to a learning rate of 2.0×10^{-5} , a batch size of 40, and 10 epochs. These parameters were set based on the settings used in the original study [11]. $J = 100$ captions were generated for each image, based on the probability distribution of the language model. In the attribute scoring module, 292 attributes common to FACAD170K and DeepFashion were used as the attribute set K to be evaluated. Caption scoring based on top attribute comparisons in the caption scoring module uses the top $n = 9$ attributes. The values were determined based on preliminary experiments that compared the estimated and correct attributes for different values of n .

The model was evaluated by comparing the inferred results of the model against FACAD170K and DeepFashion with correct labels. The evaluation metrics are as follows. The set of attributes annotated for a garment image x_i is the overall attribute set K_i^{GT} , and the set of attributes with only one garment image is the differential attribute set D_i^{GT} . This is expressed in (11) and (12).

$$D_1^{GT} = K_1^{GT} \setminus K_2^{GT} \quad (11)$$

$$D_2^{GT} = K_2^{GT} \setminus K_1^{GT} \quad (12)$$

Let $K_{y'_i}$ be the attribute set contained in the caption y'_i . The precision, Recall, and F1 scores were calculated between $K_{y'_i}$ and the differential attribute set D_i^{GT} to assess the degree of description of the attributes that differed between garments. Similar indices were calculated between $K_{y'_i}$ and the overall attribute set K_i^{GT} as supplementary indices to assess the degree of description of the attributes in each garment image. Larger values of these indices are preferable. The evaluation was performed on 10,000 pairs, and the average value of each evaluation indicator was calculated.

2) *Results:* The evaluation results for the captions generated by the proposed method in FACAD170K and DeepFashion are listed in Tables VII and VIII. A comparison of the results across datasets shows that the evaluation values for FACAD170K are higher than those of DeepFashion for all indicators. This is because the image-caption generation model ClipCap was trained on the FACAD170K data; consequently, the attribute information of FACAD170K was more appropriately reflected in the captions. For attribute scoring methods, frequency-of-occurrence-based attribute scoring tends to perform better than attribute estimation-based attribute scoring on both datasets. In particular, FACAD170K outperformed the attribute scoring based on attribute estimation for all evaluation indicators. Regarding caption scoring methods, caption scoring based on relative score addition outperformed caption scoring based on top-attribute comparisons for all evaluation indices in both datasets. These results indicate that under the experimental conditions of this study, the combination of attribute scoring based on the frequency of occurrence and caption scoring based on relative score addition is the most effective.

B. Prompt-based VLM Method

As stated in the previous section, three candidates exist for the VLM: OpenAI GPT-4o, Anthropic Claude 3.5 Sonnet, and Google Gemini Pro 1.5. Their effectiveness in generating a caption highlighting differences between clothing items is compared with the previously discussed prompts. In addition to that, the effectiveness and necessity of prompting techniques of few-shot examples and chain-of-thought reasoning are verified.

1) *Methods:* In this experiment, one of the authors annotated qualitative evaluations on generated captions for 15 pairs of clothing images. Since the annotations are time-intensive, experiments are systematically conducted on limited combinations of VLMs and prompting techniques.

Generated captions were evaluated by the following annotations and text length. A five-point Likert scale is utilized for concreteness and accuracy for the following clothing attributes. Concreteness is annotated based on whether the attribute is explained in the caption, while accuracy is given by whether the description of the attribute is correct.

- Design, Silhouette, and Details
- Material
- Color, Pattern, and Print
- Wearing Scene
- Comparison with another clothing item

for the derivative attributes, only concreteness is annotated because of their subjective nature.

- Impression
- Effect
- Wearing Scene
- Coordination and Styling
- People recommended for

In the preliminary experiment, the zero-shot setting does not produce captions containing all attributes specified in the prompt with any of the VLMs. Therefore, each VLM is combined with few-shot examples. Three VLMs were compared with few-shot examples without the chain-of-thought reasoning. The effectiveness of chain-of-thought reasoning is compared with the best-performed VLM.

2) *Result:* Of the captions generated by three VLMs with few-shot examples, the average text length did not significantly exceed the target of 100 words with all VLMs. Claude 3.5 Sonnet surpassed other models in all attributes for average concreteness and accuracy, describing design, color, impression, and effect for almost all pairs. The performance on material, coordination, and recommended people were worse than on other attributes. Gemini 1.5 Pro performed inferiorly, particularly in terms of descriptions of materials and coordination.

Introducing chain-of-thought reasoning improved the performance of Claude 3.5 Sonnet in terms of materials, coordination, and recommended people while preserving performance in other attributes and text length compliance. The prediction result of attributes as the first step was accurate for all pairs, possibly improving the captions.

From the experiment results, Claude 3.5 Sonnet with few-shot examples and chain-of-thought reasoning is adopted in the prompt-based VLM.

VI. QUALITATIVE EVALUATION OF GENERATED CAPTIONS

A. Objectives

This experiment evaluates the effectiveness of the proposed caption-generation methods by assessing the accuracy of attribute description, the clearness of explanation for differences between pairs, and the usefulness in clothing item comparison.

B. Methods

In this experiment, the captions generated by the proposed methods were presented to a group of subjects for evaluation. The subject group comprised ten male and female subjects in

TABLE VII
RESULTS IN FACAD170K.

Attribute Scoring	Caption Scoring	Differential Attributes			Overall Attributes		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
Attribute Estimation	Comparison of Top Attributes	0.145	0.171	0.144	0.198	0.174	0.173
	Relative Score Addition	0.157	0.223	0.172	0.212	0.225	0.206
Frequency of Occurrence	Comparison of Top Attributes	0.204	0.324	0.236	0.248	0.294	0.258
	Relative Score Addition	0.214	0.369	0.256	0.274	0.353	0.297

TABLE VIII
RESULTS IN DEEPFASHION.

Attribute Scoring	Caption Scoring	Differential Attributes			Overall Attributes		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
Attribute Estimation	Comparison of Top Attributes	0.051	0.089	0.058	0.077	0.091	0.077
	Relative Score Addition	0.070	0.136	0.084	0.123	0.164	0.131
Frequency of Occurrence	Comparison of Top Attributes	0.057	0.139	0.075	0.088	0.143	0.104
	Relative Score Addition	0.059	0.156	0.080	0.096	0.171	0.118

TABLE IX
SET QUESTIONS AND OPTIONS.

No.	Question
Q1	Do you think the description of the attributes of Garment A/B is specific and accurate?
Q2	Do you think the description of the derivation based on the attributes of Garment A/B is specific and accurate?
Q3	Do you think that the attributes and derivatives unique to Garment A/B are described specifically and accurately?
Q4	Do you think a clear comparison is being made with Garment B/A in the caption to Garment A/B?
Q5	Do you think the two captions help you compare garments when you are choosing one to buy?

their 20s. Five pairs of clothing images were prepared. Two proposed algorithms are applied for each pair, and two pairs of captions are obtained. The clothing images and caption pairs were presented to the subjects without specifying the generation method and were evaluated.

Two examples of the presented pairs of clothing images and generated captions are shown in Figure 4. The other three pairs are provided in the supplementary. The pairs consist of highly similar clothing items based on preliminary experiments indicating that captions are most needed when distinguishing between highly similar clothing items. Each item of the pair is referred to as Garment A or B in the captions and questionnaire.

Table IX shows the questions and options set. To make the terms of attributes, derivatives, and these unique to one garment and clear comparison explicit to the subjects, the caption of Figure 2 and each corresponding part of the text was shown to the subjects in advance. Q1 and Q2 were designed to assess how accurately the caption represented garment attributes. Q3 and Q4 assessed how well the captions described the differences between items. Furthermore, Q5 was established to test the usefulness of the caption pairs provided for comparing garments. Q1 to Q4 were answered for each caption of pairs, whereas Q5 was asked for each caption pair. A five-point Likert scale was used to answer each question as

follows.

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

In addition, the subjects were asked to explain their answers to Q5 and any erroneous features and features not described in the caption.


Wilcoxon's signed-rank test was used as the test method. This test checked whether the answers to each question for the caption generated by prompt-based VLM were significantly biased from neutral and whether there were significant differences between the caption generation methods. The significance level was set at 5%. Furthermore, Bonferroni correction was applied to account for the effects of multiple tests.

C. Results

The proportions of the responses to each question are shown in Figure 5. For the captions and pairs generated by the caption pair selection method, approximately half of the respondents positively answered 'strongly agree' or 'agree' in Q1. In the rest of the questions, the percentage of favorable responses was well below half. In particular, more than 95% of the answers were 'strongly disagree' or 'disagree' in Q4. For those generated by the prompt-based VLM method, more than 80% of the respondents answered 'strongly agree' or 'agree' in all questions, and more than half answered 'strongly agree.' The percentage of negative answers was relatively high in Q5, where about 10% of the respondents answered 'strongly disagree' or 'disagree'. The p-value of the Wilcoxon signed-rank test was below 0.5%, the significance level after adjusting with Bonferroni correction, for all comparisons between the caption pair selection and the prompt-based VLM method and bias from neutral for the prompt-based VLM method.

We obtained the following explanations from respondents for negative answers in Q5 towards the caption by prompt-based VLM.

1A




Caption Pair Selection

Breathable slubbed tech modal and piqu knit ensure performance comfort in a standby polo with easy appeal.

Prompt-based VLM

This light blue polo shirt offers a classic, versatile style perfect for various occasions. Its button-down collar and 3-button placket provide a refined touch, while the soft, lightweight knit fabric ensures all-day comfort. Unlike Garment B's bold stripes, this polo's solid heathered color allows for easy pairing with multiple bottoms. Ideal for those who prefer a timeless, understated look, this shirt transitions seamlessly from casual office settings to weekend outings, making it a wardrobe essential for the modern man seeking both style and versatility.

1B



Caption Pair Selection


Breathable slubbed tech modal and piqu knit ensure performance comfort in a standby polo with easy appeal.

Prompt-based VLM

Make a statement with this blue and white striped camp collar shirt. Its relaxed fit and full button-front closure create a breezy, laid-back silhouette perfect for summer events. Unlike Garment A's more conservative polo style, this shirt features a trendy camp collar and eye-catching vertical stripes, ideal for those who enjoy expressing their personality through fashion. The lightweight woven fabric keeps you cool during beach vacations or outdoor parties. Recommended for the fashion-forward individual who embraces bold patterns and a vacation-inspired aesthetic in their wardrobe.

(a) Pair 1.

2A




Caption Pair Selection

Outfitted with elasticized drawstring waist cargo pocket and utilitarian inspired patch pocket these crew pants are the ultimate in rugged utility.

Prompt-based VLM

These olive green straight-leg pants offer a perfect blend of casual comfort and smart style. Unlike Garment B's relaxed jogger design, these feature a more structured silhouette with a button closure and belt loops, making them versatile for both weekend outings and smart-casual work environments. The tapered ankle with elastic cuffs provides a modern touch, elongating the legs while maintaining comfort. Ideal for those seeking a polished casual look, these pants pair effortlessly with t-shirts or button-ups, offering greater versatility than the sportier alternative.

2B



Caption Pair Selection

Soft and stretchy cotton blend fleece mean easy comfort in comfort joggers with an elastic drawstring waistband.

Prompt-based VLM

Embrace comfort with these charcoal grey jogger-style pants, designed for ultimate relaxation and ease of movement. In contrast to Garment A's more structured design, these pants feature a soft, stretchy fabric with a drawstring waist, perfect for athletic activities or lounging. The heathered texture adds visual interest, while the tapered leg and cuffed ankles provide a trendy silhouette. Ideal for those prioritizing comfort and a sporty aesthetic, these pants are best paired with t-shirts and hoodies for a laid-back look that Garment A's more tailored style doesn't offer.

(b) Pair 2.

Figure 4. Examples of presented clothing images pair and captions shown to participants.

Pair 1 It is difficult to read the intent compared to the reference captions, whether it is a comparison of clothing features or impressions.

Pair 4 Items are so different that I don't really feel the need to refer to the caption.


Pair 5 The caption for Garment A describes it as "Unlike Garment B, this jacket is ideal for mild weather" but the caption for Garment B states that it is "While less suitable for rainy conditions than Garment A" giving the impression that the opposite is true

Pair 5 There are some errors in the characteristics, and I feel a discrepancy with the impression I actually have.

Pair 5 It is hard to tell what they are comparing. I felt like there was a mix of functional and visual comparisons compared to the reference captions.

In most explanations for negative answers to the caption pair selection method, it is pointed out that the captions are short, and described attributes are easily understandable only from the images without reading the caption.

3A




Caption Pair Selection

With its bubbly print and puffed sleeve this polka dot blouse is a delightful way to add a fun and playful air to your ootd.

Prompt-based VLM

This charming mint green blouse features a playful polka dot pattern and feminine details like a tie-front neckline, gathered shoulders, and ruffled cuffs. The light, flowy fabric and short sleeves make it perfect for casual outings or creative office environments. Unlike Garment B's solid color and button-front design, this blouse offers a more whimsical, retro-inspired look. Ideal for those who love incorporating patterns and vintage-inspired pieces into their wardrobe, it pairs beautifully with high-waisted jeans or skirts for a balanced silhouette.

3B



Caption Pair Selection


Embroidered trim accentuates the v neck of this dreamy striped top cut with a split neckline that play up the stripey sophistication.

Prompt-based VLM

This elegant turquoise blouse combines modern simplicity with feminine touches. Featuring a V-neckline, button-front closure, and ruffled cuffs, it offers versatility that Garment A's tie-front and polka dot design can't match. The lightweight, slightly textured fabric in a solid color makes it perfect for office wear or semi-formal events. Unlike Garment A's more whimsical style, this blouse provides a streamlined look that's easy to dress up or down. Ideal for those seeking a versatile piece, it layers beautifully under blazers or cardigans for a polished, professional appearance.

(c) Pair 3.

4A




Caption Pair Selection

Cutoff elbow release provide a fresh twist on a belt bag skirt made from slick leather and one-piece on an ultrafine faux suede sole.

Prompt-based VLM

Make a bold statement with this sleek black faux leather midi skirt. Its A-line silhouette and button-down front create a striking, edgy look that's perfect for evening events or fashion-forward workplaces. Unlike Garment B's relaxed fit, this skirt offers a more structured, form-fitting style that elongates the legs and defines the waist. The smooth, glossy texture adds a touch of sophistication, making it ideal for those who want to stand out with a modern, daring ensemble. Pair with fitted tops for a sleek silhouette that's sure to turn heads.

4B



Caption Pair Selection

Shapely pleated and textured in cotton candy stripe that fade at the waist this shorts and tiered midi is the perfect complement to your 9 to 5.

Prompt-based VLM

Embrace effortless elegance with this navy blue striped wrap skirt. Its asymmetrical design and pleated texture offer a softer, more feminine alternative to Garment A's structured leather look. The lightweight fabric and tie waist provide comfort and versatility, making it perfect for office wear or casual outings. The vertical stripes elongate the figure, while the wrap style allows for easy movement. Unlike Garment A's edgy appeal, this skirt exudes a relaxed, summery vibe that's ideal for those who prefer a more fluid, adaptable piece in their wardrobe. Style with a simple top for a chic, put-together look.

(d) Pair 4.

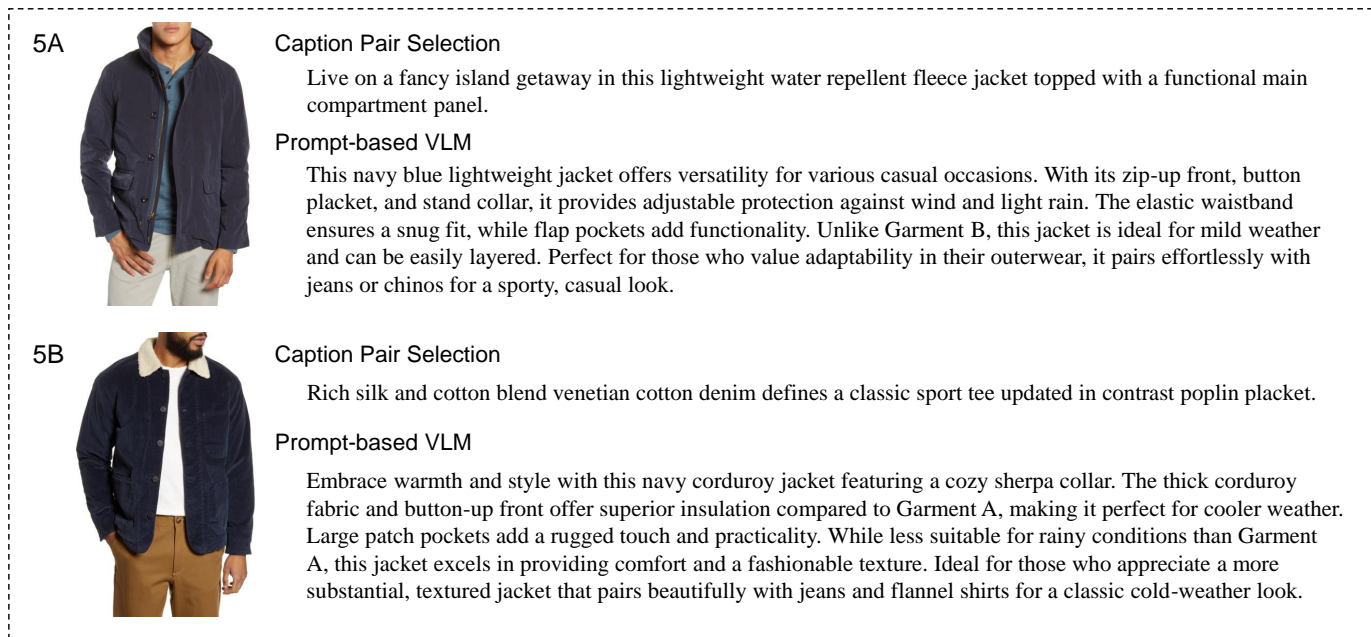
Figure 4. Examples of presented clothing images pair and captions shown to participants.

D. Discussions

The overall result indicates that the caption pairs generated by the prompt-based VLM method are preferred to those generated by the caption pair selection method. From the respondents' explanation, the utilized existing caption generator does not cover all the attributes required for the captions that highlight differences. The fact that the proportion of positive answers in Q2 is lower than in Q1 for caption pair selection implies that the caption generator only describes clothing attributes such as design and material and fails to explain

derivatives like impression and coordination. Additionally, a high proportion of negative answers in Q4 shows the limitation of caption pair selection in that the existing generator cannot generate captions that explicitly describe differences, and thus, such captions cannot be selected. These weaknesses can be covered by fine-tuning, although it requires a dataset of captions that contains derivatives. On the other hand, the flexibility of the state-of-the-art VLM can address such an issue by in-context learning with only a few examples.

The negative answers in Q5 for prompt-based VLM indicate



(c) Pair 5.

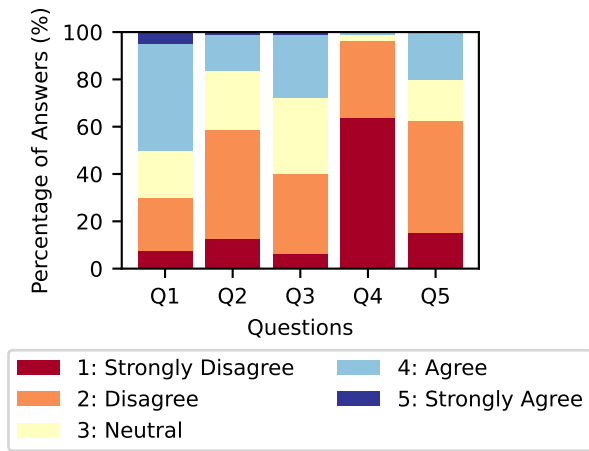
Figure 4. Examples of presented clothing images pair and captions shown to participants.

room for improvement in this approach. One of the reasons is the lack of logical consistency in the generated captions. For example, a further interview for a negative answer in Pair 5 revealed that the respondent felt uncomfortable with the phrase “The thick corduroy fabric and button-up front offer superior insulation compared to Garment A”. It is natural to think that insulation is introduced only by corduroy fabric, not by button-up front. Another example is that one of the respondents found that the part of the caption for 1A, “Unlike Garment B’s bold stripes, this polo’s solid heathered color allows for easy pairing with multiple bottoms.” describes a clothing feature, while the corresponding description for 1B, “Unlike Garment A’s more conservative polo style, this shirt features a trendy camp collar and eye-catching vertical stripes, ideal for those who enjoy expressing their personality through fashion.” shows an impression, giving a misleading feel. The caption for 1A implicitly includes a nuance that 1A is less eye-catching and thus can be combined with various bottom items, which can be explicitly stated for a clearer caption. These logical issues could potentially be resolved by extending the chain-of-thought reasoning so that relations between attributes are inferred. Another reason can be the subjective nature of derivative attributes since one respondent found a discrepancy between the description and her or his impression for pair 5 while others necessarily did not. This issue suggests that personalization can be required to describe such attributes.

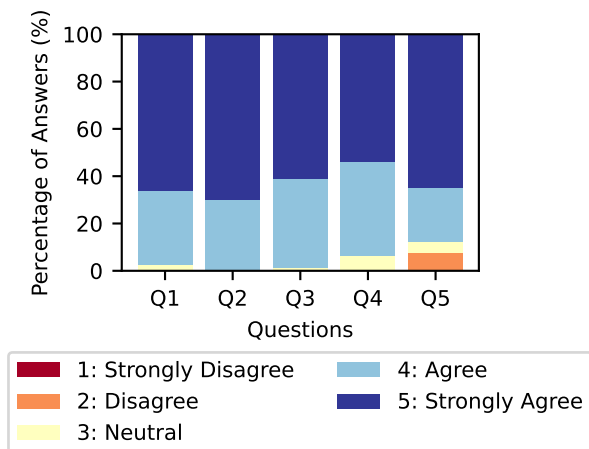
VII. CONCLUSION AND FUTURE WORK

This study proposed and evaluated a caption-generation method that highlights the differences between pairs of garment images to provide helpful information for consumers when comparing products. The content that should be included

in the captions for the clothing item comparison is assessed from lifestyle magazine articles. The method to generate the captions without preparing a task-specific dataset is discussed, and two methods are proposed, namely caption pair selection and prompt-based VLM. In the caption pair selection method, two garment images are input independently into an image caption generator to generate multiple captions. Attribute scores are then calculated for each image. A caption score is then calculated for each caption in the multiple captions generated for each image using the attribute scores. Finally, the captions are selected and output based on caption scores. Automatic evaluation experiments were conducted on attribute scoring and caption scoring, focusing on accurately describing the features of a single garment and the differences between garments. In the prompt-based VLM method, two images are given to the VLM, and a pair of captions are generated simultaneously. The prompt is designed based on the assessment to describe clothing attributes and item comparisons. Additionally, prompting techniques for few-shot examples and chain-of-thought reasoning are utilized. Since there are multiple approaches and VLMs to implement these methods, preliminary experiments are conducted. Methods employing attribute scoring based on the frequency of occurrence and caption scoring based on relative score addition were rated highly. Attribute scoring based on frequency of occurrence uses the frequency of an attribute’s occurrence in the caption as the attribute score, whereas caption scoring based on relative score addition calculates the relative value of the attribute score and adds it to the number of attributes that appear. Furthermore, captions generated by a combination of methods that received high ratings in the automatic evaluation experiment were presented to the subjects, and a qualitative evaluation of their useful-



(a) Caption Pair Selection.



(b) Prompt-based VLM.

Figure 5. Percentage of answers to each question for captions generated by proposed methods.

ness was conducted. Multiple state-of-the-art VLMs publicly available through APIs are evaluated by annotating generated captions using several criteria determined by the assessment of the item comparison article. As a result, Claude 3.5 Sonnet is selected, and the effectiveness of chain-of-thought reasoning by estimating clothing attributes is verified. The quantitative evaluation with a questionnaire revealed that the prompt-based VLM generates captions containing the required content for comparison and provides helpful information for comparing two garments with the flexibility of the state-of-the-art VLM. Furthermore, it is confirmed that the proposed method has room for improvement due to the need for more logical consistency and subjectivity of the clothing attributes.

The proposed method can only specify two garment images as input images. We plan to extend this approach to handle more than three garment images to meet consumer garment comparison needs better. Based on the assessment of clothing item comparison articles, captions generated for each item can be concatenated to provide information for comparison. The prompt-based VLM method can be easily extended for

multiple clothing items, provided that the sequence length for inputs and outputs of VLM is sufficient.

REFERENCES

- [1] A. Kohei, Y. Soichiro, Y. Tomohisa, and K. Hidenori, "Generation of Captions Highlighting the Differences between a Clothing Image Pair with Attribute Prediction," in *INTELLI 2024, The Thirteenth International Conference on Intelligent Systems and Applications*, 2024, pp. 7–16.
- [2] M. R. Solomon, *Consumer Behavior: Buying, Having, and Being*. Pearson, 2020.
- [3] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, Springer, 2014, pp. 740–755.
- [4] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [5] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.
- [6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.
- [7] S. Ioffe, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv preprint arXiv:1502.03167*, 2015.
- [8] A. Graves and A. Graves, "Long Short-Term Memory," *Supervised Sequence Labelling with Recurrent Neural Networks*, pp. 37–45, 2012.
- [9] K. Xu *et al.*, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 2048–2057.
- [10] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical Sequence Training for Image Captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7008–7024.
- [11] R. Mokady, A. Hertz, and A. H. Bermano, "ClipCap: CLIP Prefix for Image Captioning," *arXiv preprint arXiv:2111.09734*, 2021.
- [12] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, 2021, pp. 8748–8763.
- [13] A. Radford *et al.*, "Language Models are Unsupervised Multitask Learners," 2019.

- [14] P. Wang *et al.*, “OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework,” *CoRR*, vol. abs/2202.03052, 2022.
- [15] L. Ouyang *et al.*, “Training Language Models to Follow Instructions with Human Feedback,” in *Advances in Neural Information Processing Systems*, S. Koyejo *et al.*, Eds., vol. 35, Curran Associates, Inc., 2022, pp. 27 730–27 744.
- [16] J. Achiam *et al.*, “GPT-4 Technical Report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [17] N. Wang, J. Xie, J. Wu, M. Jia, and L. Li, “Controllable Image Captioning via Prompting,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 2617–2625.
- [18] Y. Ge *et al.*, “Visual Fact Checker: Enabling High-Fidelity Detailed Caption Generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 033–14 042.
- [19] H. Jhamtani and T. Berg-Kirkpatrick, “Learning to Describe Differences Between Pairs of Similar Images,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 4024–4034.
- [20] D. H. Park, T. Darrell, and A. Rohrbach, “Robust Change Captioning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4624–4633.
- [21] J. Wang, W. Xu, Q. Wang, and A. B. Chan, “Group-based Distinctive Image Captioning with Memory Attention,” in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5020–5028.
- [22] Y. Mao *et al.*, “Rethinking the Reference-based Distinctive Image Captioning,” in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 4374–4384.
- [23] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, “Where to Buy It: Matching Street Clothing Photos in Online Shops,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3343–3351.
- [24] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [25] J. Deng *et al.*, “ImageNet: A Large-Scale Hierarchical Image Database,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [26] J. Huang, R. S. Feris, Q. Chen, and S. Yan, “Cross-Domain Image Retrieval with a Dual Attribute-Aware Ranking Network,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1062–1070.
- [27] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, “Deep-Fashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1096–1104.
- [28] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, “Deep Convolutional Ranking for Multilabel Image Annotation,” *arXiv preprint arXiv:1312.4894*, 2013.
- [29] A. Sonoda, “Apparel EC Saito Ni Okeru Setsumei Bun Jidou Seisei (Automatic Generation of Descriptions in Apparel E-Commerce Sites),” in *Proceedings of the Japan Society of Management Information National Conference, 2018 Autumn*, Japan Society of Management Information, 2018, pp. 125–127.
- [30] X. Yang *et al.*, “Fashion Captioning: Towards Generating Accurate Descriptions with Semantic Rewards,” in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, Springer, 2020, pp. 1–17.
- [31] C. Cai, K.-H. Yap, and S. Wang, “Attribute Conditioned Fashion Image Captioning,” in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 1921–1925.
- [32] Q. Chen *et al.*, “Fashion-GPT: Integrating LLMs with Fashion Retrieval System,” in *Proceedings of the 1st Workshop on Large Generative Models Meet Multimodal Applications*, 2023, pp. 69–78.
- [33] Y. Ding *et al.*, “FashionReGen: LLM-Empowered Fashion Report Generation,” in *Companion Proceedings of the ACM on Web Conference 2024*, 2024, pp. 991–994.
- [34] A. Maronikolakis, A. P. Ramallo, W. Cheng, and T. Kober, “What Should I Wear to a Party in a Greek taverna? Evaluation for Conversational Agents in the Fashion Domain,” *arXiv preprint arXiv:2408.08907*, 2024.
- [35] F. Carlsson, P. Eisen, F. Rekathati, and M. Sahlgren, “Cross-Lingual and Multilingual CLIP,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 6848–6854.
- [36] N. Rostamzadeh *et al.*, “Fashion-Gen: The Generative Fashion Dataset and Challenge,” *arXiv preprint arXiv:1806.08317*, 2018.
- [37] S. Guo *et al.*, “The iMaterialist Fashion Attribute Dataset,” in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 3113–3116. DOI: 10.1109/ICCVW.2019.00377.
- [38] OpenAI. “Hello GPT-4o,” Accessed: 2024-12-12. [Online]. Available: <https://openai.com/index/hello-gpt-4o/>.
- [39] Anthropic. “Introducing Claude 3.5 Sonnet,” Accessed: 2024-12-12. [Online]. Available: <https://www.anthropic.com/news/claude-3-5-sonnet>.
- [40] Gemini Team, Google *et al.*, “Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context,” *arXiv preprint arXiv:2403.05530*, 2024.
- [41] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual Instruction Tuning,” in *NeurIPS*, 2023.

Day-ahead Forecasting Electricity Spot Prices in Norway with ARIMA, XGBoost, and LSTM Models

1st Markus Wiik Jensen

School of Economics,
Innovation and Technology
Kristiania University College
Oslo, Norway

email:maje012@student.kristiania.no

2nd Huamin Ren

School of Economics,
Innovation and Technology
Kristiania University College
Oslo, Norway

email:huamin.ren@kristiania.no

3rd Andrii Shalaginov

School of Economics,
Innovation and Technology
Kristiania University College
Oslo, Norway

email:andrii.shalaginov@kristiania.no

Abstract—This paper comprehensively explores univariate and multivariate forecasting models for the Norwegian Elspot markets. As a leading renewable energy supplier with a high reliance on hydropower, Norway offers valuable insights into balancing renewable sources. The volatility of its electricity market, influenced by broader European trends, underscores the need for accurate forecasting. Day-ahead electricity price forecasts from the Elspot market are crucial for electricity producers and market operators, informing supply bids and dispatch schedules. This research includes experiments with advanced forecasting methods, combining machine learning and time series analysis to improve accuracy. We compare three models—ARIMA, XGBoost, and LSTM—across Norway’s six Elspot markets. LSTM outperforms the other models in three specific zones, demonstrating its superior predictive performance. Future research will focus on enhancing model generalization.

Index Terms— Green Energy; Electricity Price Forecasting; Elspot prices; XGBoost; LSTM.

I. INTRODUCTION

Electricity and energy are integral to modern society, driving economic growth, technological advancement, and overall quality of life. Energy consumption, closely linked to factors such as wealth, health, and infrastructure, has been steadily increasing due to population growth, industrialization, and technological development. As the world transitions away from fossil fuels and seeks sustainable solutions, understanding and forecasting energy markets becomes crucial. Accurately forecasting market trends and price fluctuations is of paramount significance for a diverse range of stakeholders, including investors, businesses, and policymakers [10] [14] [16] [31]. The Norwegian electricity markets, characterized by deregulation and high renewable integration, present unique research opportunities. Recent market disruptions, marked by volatile prices and increased uncertainty, underscore the need for advanced forecasting techniques. This research aims to enhance understanding of Norway’s electricity markets by investigating key price drivers and evaluating electricity price forecasting (EPF) methods. Such predictions are crucial for electricity producers, consumers, and market operators to effectively plan their production, consumption and trading activities [3].

The NordPool spot (Elspot) market is a day-ahead market, where the price of power is determined by supply and demand. Such spot prices are the actual prices for electricity for the next day, and will be set at NordPool Elspot. Our primary focus is on day-ahead price forecasting using known spot prices. This forecasting directly informs bidding strategies for the upcoming day [19]. Due to the distinct characteristics of electricity markets, each forecasting challenge is unique across different markets and necessitates bespoke model developments [24]. We propose a framework for evaluating forecasting methods for all six Elspot markets of Norway while comparing three different numerical approaches to the problem of extrapolating prices in both univariate and multivariate configurations, facilitating the identification of region-specific models and model configurations. By examining the factors influencing price dynamics and comparing various forecasting methodologies, this study seeks to improve predictive accuracy and interpretability. The findings will provide a framework for future research and support decision-making in modeling and market analysis. In Section II, we dive into electricity markets and existing literature on EPF. Section III presents the methodologies employed. Section IV discusses the conducted experiments, and in Section V we conclude with an analysis of the obtained results.



Figure 1. Hydropower reservoir.

II. BACKGROUND AND LITERATURE STUDIES

In this section we review the various market mechanics characterizing electricity markets and existing literature concerning EPF.

A. Background

Electricity is produced only moments before consumption, so unlike other commodities, electricity must be balanced between production and consumption at all times [17]. In a deregulated market environment, determining the unconstrained Market Clearing Price (MCP), commonly referred to as the spot price of an electricity pool typically involves the following steps:

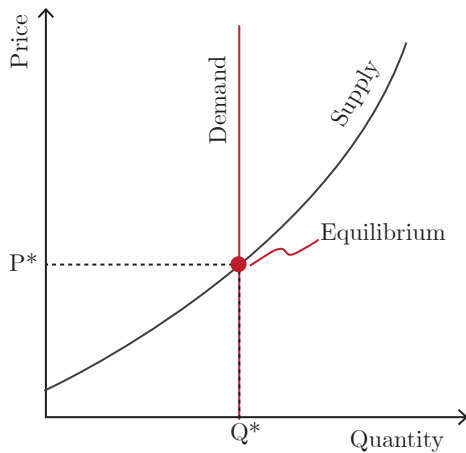


Figure 2. Equilibrium curve to determine the MCP of a bidding-pool.

- Generating companies bid prices for supplying energy, creating a supply curve.
- The demand curve may be set at a value derived from a forecast of the load due to short-term inelasticity for demand of electricity, resulting in a vertical line at the forecasted load value.
- Spot price is found where supply and demand curves intersect, signifying the market equilibrium.

The spot price is set at the equilibrium between supply and demand as seen in Figure 2 for each hour of the following day after accounting for the bids received within the deadline as illustrated in Figure 3 [14].

Like many goods and services, electricity demand exhibits daily, weekly, and seasonal fluctuations. Consumption typically peaks during late afternoon and early evening when

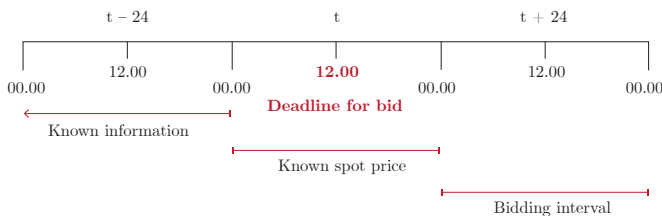


Figure 3. Deadline for bids in the Elspot markets.

people return home from work and school, activating their lighting and appliances. This period, known as "peak demand," necessitates increased electricity generation to meet the higher demand. Demand patterns also vary by location, influenced by local weather conditions and regional consumption behaviors [31]. For instance, in warmer climates during summer, electricity demand rises due to increased use of air conditioning. Conversely, in colder regions during winter, higher demand is driven by heating requirements. Understanding these demand factors is crucial for utilities and policymakers to ensure a reliable and sustainable electricity supply [14]. Electricity generation encompasses various methods, including coal, natural gas, nuclear power, and renewables such as wind, solar, and hydro power. Due to the non-storability of electricity and the need for load management, generators must continuously adjust their output to balance supply and demand. This task is particularly challenging for renewable sources like wind and solar power that are inherently intermittent and uncontrollable. Wind and solar energy production cannot be precisely controlled and is dependent on current weather conditions, resulting in variable output. In contrast, nuclear power provides a stable, continuous supply but lacks the flexibility to adjust output quickly in response to demand changes. Technologies capable of responding to rapid fluctuations, such as flexible hydro-power and liquid natural gas (LNG) generation, are essential for maintaining grid stability. Additionally, the value of storable production sources, such as fossil fuels and hydro, is influenced by their convenience yield—an extra benefit from holding the commodity beyond potential financial gains from its sale.

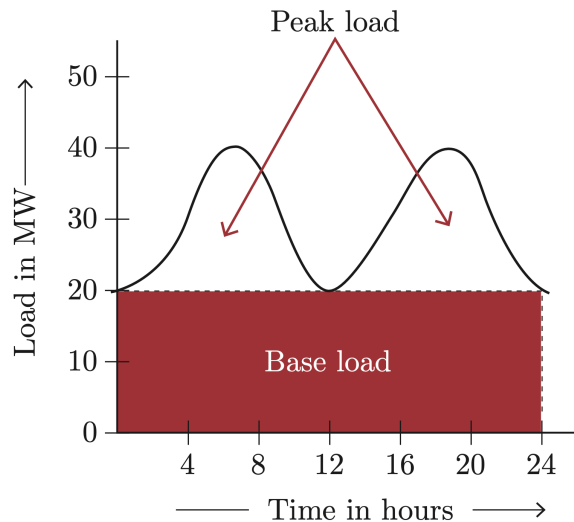


Figure 4. Base-load vs. Peak-load.

Mechanics such as electricity market and pricing, electricity production and consumption are crucial to understand the complexity of EPF. The electricity market is influenced by a multitude of factors, including supply and demand dynamics, changing industrial and household consumption,

multiple seasonality, weather conditions, regulatory policies, fuel prices, the integration of renewable energy sources, and the rapid diffusion of price-anomalies [3] [10] [14] [16] [31]. Understanding the key drivers of price movements aids in feature selection for predictive models. For instance, if weather patterns or economic variables significantly affect prices, incorporating these into a model may improve accuracy. The choice of methodology should also consider the nature of price drivers, as incorporating these considerations guides model selection. Furthermore, accurate price forecasts coupled with an understanding of their drivers provide valuable market insights.

A time series is defined as a series of data points indexed in time order [34]. Commonly expressed as:

$$X = X_{t=1}^{\infty} = (X_1, X_2, \dots) \quad (1)$$

where X_t denotes the observation at time t , and the sequence of observations is indexed by t ranging from 1 to infinity.

Accurately extrapolating the future electricity prices poses unique challenges due to several constraints imposed by time order. Some of these constraints include look-ahead bias, stationarity, auto-correlation, seasonality, trend and noise. Time-series data, characterized by sequential observations over time, requires specialized methodologies that can capture temporal dependencies and patterns. Time series forecasting (TSF) attempts to predict future outcomes based on historical context and has direct applications to many domains, including science, policy and business. Because TSF is based on historical data it can be useful for planning future actions based on previous actions, by measuring the statistical correlations between variables over time to predict the future it is possible to also explore meaningful patterns in the data that would otherwise remain inconceivable.

B. Literature Studies

In the domain of EPF, selecting appropriate input variables, historical data duration, and modelling techniques is crucial. Most efforts that focus on forecasting day-ahead prices typically experiment with an inference horizon of 1-4 weeks [4] [5] [12] [13] [15] [19] [20] [24] [25] [28] [32] [35]. Historical data spanning at least a year is commonly employed to capture yearly seasonality [4] [15] [20] [26] [35]. Input variables encompass a range of factors, including past prices [4] [5] [7] [8] [11] [12] [13] [15] [18] [19] [20] [23]- [29] [32] [35], system loads [15] [19] [23] [25]- [28] [32], weather variables [7] [15] [20] [26] [33], fuel costs [5] [7] [21] and sector indices [30]. Preprocessing and data transformations are essential to handle missing values and outliers that can affect model performance. Techniques like normalization [7] [8] [32], decomposition [8] [12] [20] [25] [27] [35], and differentiation [13] are used to improve data quality and model accuracy. Statistical models, such as econometric methods, like Linear Regression [15] [23] [25] [33] and Auto-Regressive models [5] [12] [13] [15] [18] [20] [32] [35], offer interpretability and insights into correlations. Algorithmic models like Deep Learning (DL) [8]

[15] [18] [19] [21] [23]- [27] and Ensemble models [5] capture complex and nonlinear patterns.

As highlighted in numerous studies, the process of building a forecasting model involves decisions on input selection, forecasting horizons, preprocessing and feature engineering techniques, model choice, parameter estimation, and accuracy evaluation. However, guidelines for navigating these complexities are limited, with much variation in reported approaches. Given the specific nature of EPF, establishing baselines and ensuring rigorous reporting is critical for advancing research in this field.

The process of determining critical design decisions vary, Amjadi and Hemmati [3] emphasize that most input-variable selections are based on forecaster heuristics rather than a systematic approach, while Aggarwal et al. [2] note that advancements in technology, market optimization, and data availability continue to alter the landscape of optimal variable selection. A universal set of price drivers is unlikely to emerge, given the diverse nature of electricity markets. Despite the growing number of studies on EPF, many lack transparency and statistical rigor, making it difficult to compare different approaches. Studies focusing on advanced statistical techniques often compare these only to basic machine learning (ML) methods, while ML-based studies typically contrast against simple statistical techniques, further complicating cross-study comparisons. Major review publications have pointed out that inconsistent datasets, implementations, error measures, and problem definitions exacerbate this issue, making it hard to assess the transferability of findings to other markets or future developments [2] [3] [24] [32]. The failure to control for issues like data contamination and look-ahead bias is common, as many studies do not specify details such as test-train splits, input variables, or data transformations. Lago et al. [24] stress the importance of ensuring that the test dataset is always the last segment of the full dataset, with no overlap with training data. Moreover, a significant number of studies neglect to benchmark new methods against simpler, well-established models such as naive heuristics, as they are crucial for evaluating the true generalization performance of complex models. The lack of such baselines can lead to spurious conclusions about model performance, even in otherwise well-conducted research. Future studies should rigorously incorporate these practices to enhance reproducibility, validity, and the significance of results.

III. PROPOSED METHODOLOGY

In this research, the approach begins with selecting a baseline method that is heuristic-based. Building upon this baseline, the study conducts an empirical-driven progression to develop previously proven forecasting models in both univariate and multivariate configurations. Three distinct approaches are explored: a econometric method, an algorithmic ensemble approach, and a deep learning (DL) approach. This methodology is designed to ensure objectivity and standardization in the evaluation process. Given the unique and inconsistent

nature of electricity markets, EPF challenges vary significantly across locations and time frames, rendering cross-study evaluations potentially misleading and universal benchmarks logically unsound for this domain. Therefore, the methodology involves systematic steps, including literature review of related work, data collection and preparation, model development and rigorous testing against real world outcomes. Models are trained, validated, and tested in both univariate and multivariate configurations, enabling comparisons of the added value of incorporating exogenous variables. The expectation is that multivariate models should outperform univariate models, that rely solely on price data, to justify the increased complexity and computational cost. By comparing forecast results from both configurations, the study aims to shed light on the role of exogenous variables as price drivers. All the data-handling, -visualization and model-implementation and -evaluation was done using Python software.

A. Heuristic Baseline

The persistence forecast is utilized as a baseline for this study. This approach involves using the last observed value of the time series as the forecast for the corresponding day-ahead time step. In the context of day-ahead EPF, this would mean using the most recent price value as its prediction for the same hour the next day. Assuming we have a time series of electricity prices $p_t, p_{t+1}, p_{t+2}, \dots, p_{t+n}$ where t is the current time step, the persistence model predicts the current price 24 hours ahead for each time step. In the context of day-ahead EPF, the persistence model serves as a sensible baseline. While more complex modelling methods may exhibit reasonable accuracy, they must be able to generalize beyond the explicit information provided in the input data. As a baseline the heuristic provides a reference point against which, more advanced models can be evaluated, ensuring that they genuinely contribute to improved forecasting performance. We can express the persistence model in mathematical notation as follows:

$$\hat{P}_t = P_{t-24} \quad (2)$$

where \hat{P}_t denotes the predicted electricity price at time t and P_{t-24} is the observed value of the electricity price 24-time steps earlier.

B. Econometric

The ARIMA (Autoregressive Integrated Moving Average) (p,d,q) method is a popular time series forecasting technique that models the time series data as a combination of autoregressive (AR) and moving average (MA) components, with an additional differencing step to account for non-stationarity. The parameters p, d, and q are integers that represent the order of the AR, differencing, and MA components, respectively. The ARIMA(p,d,q) method can be specified by the following mathematical notation:

$$Y_t = c + \sum_{i=1}^p \Phi_i Y_{t-i} + \epsilon_t - \sum_{j=1}^q \theta_j \epsilon_{t-j} \quad (3)$$

where Y_t is the value of the time series at time t , c is a constant term, $\Phi_1, \Phi_2, \dots, \Phi_p$ are the AR coefficients, ϵ_t is the error term at time t , and $\theta_1, \theta_2, \dots, \theta_q$ are the MA coefficients.

The AR component models the current value of the time series as a linear combination of its past values, with the weights determined by the AR coefficients. The MA component models the current value of the time series as a linear combination of the past errors, with the weights determined by the MA coefficients. While ARIMA models offer benefits such as capturing auto-correlation and seasonality, providing interpretability, they have limitations in handling non-linear relationships and the assumption of stationarity. These factors should be carefully considered when applying ARIMA-type models to forecast electricity prices.

C. Algorithmic Ensemble

Extreme Gradient Boosting (XGBoost) is a popular gradient-boosting algorithm that is commonly used in machine-learning applications for both classification and regression tasks. It is an ensemble algorithm that combines multiple weak models (decision trees) to make a strong prediction. XGBoost learns from examples by building a series of decision trees. Each tree tries to correct the mistakes made by the previous trees reducing the risk of overfitting, and leading to a more accurate prediction. To further identify the most impactful variables and account for non-linear relationships between targets and inputs, XGBoost's feature gain scores are employed. This metric measures the relative contribution of each feature to the objective function, with higher scores indicating greater importance in generating accurate predictions [8]. The objective function for XGBoost can be written as:

$$\mathcal{L}(\Theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (4)$$

where Θ represents the set of model parameters, n is the number of training examples, y_i is the true value of the i -th example, \hat{y}_i is the predicted value, $l(y_i, \hat{y}_i)$ is the loss function, K is the number of weak models, f_k represents the k -th weak model, and $\Omega(f_k)$ is the regularization term.

The weak models used in XGBoost are decision trees, and can be expressed as:

$$f(x) = \sum_{t=1}^T w_t q_t(x), \quad w \in \mathbb{R}^T, \quad q : \mathbb{R}^d \rightarrow \{1, 2, \dots, T\} \quad (5)$$

where x is the input features, w is the vector of weights associated with each leaf node of the tree, T is the number of leaf nodes, and $q(x)$ is the function that maps the input features to the index of the corresponding leaf node.

D. Deep Learning

Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that is commonly used for time-series forecasting. Unlike traditional RNNs, LSTM networks are designed to overcome the problem of vanishing gradients,

making it difficult for the network to learn and remember long-term dependencies in the data. In simple terms, the LSTM network is like a specialized memory unit that can selectively remember important information from the past and use it to make predictions about the future. It achieves this by using a system of gates to control the flow of information within the network.

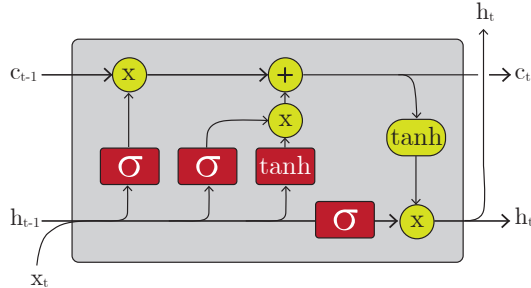


Figure 5. Long Short-Term Memory (LSTM) Network Diagram.

The LSTM network has three main types of gates as visualized in Figure 5: input gates, forget gates, and output gates. These gates allow the network to decide information that is important to keep, information to forget, and when to output its predictions [21].

IV. EXPERIMENTS AND DISCUSSION

This section covers the datasets used, the experimental setup, and the ensuing presentation and discussion of results. The research aims to clarify the methods used to forecast Norway's contemporary electricity markets by examining both modelling approaches and relevant price drivers. Studying these aspects together is beneficial, as understanding key price drivers can guide the selection of inputs that enhance model performance. For example, incorporating influential factors like weather or economic conditions can improve forecast accuracy. Additionally, the choice of modelling method depends on how price drivers interact with electricity prices, particularly in cases of non-linearity or temporal dependencies. By aligning model selection with these dynamics, we aim to create more reliable and robust forecasts. While price drivers are not the primary focus, their consideration is essential for optimizing model inputs and improving the interpretability of results. This approach promises more accurate predictions and a clearer understanding of the factors shaping Norway's electricity market.

A. Dataset and Description

Following background theory and related work, a diverse range of independent variables that are identified as potential price-drivers for the Norwegian markets were selected, from fundamental variables such as operating data and weather variables that governs production, to macro variables such as oil-prices and international or regional trade. To collect and preprocess the data, a Python environment is utilized as it is capable of handling various sources, file types, and formats. The data, including unit measures, granularity and data sources

are described in Table I. A total of six data-sets were created, each comprising time series data from one of the six bidding zones. Comparing electricity prices across different regions can provide insights into the factors driving price dynamics in a specific region and guide the development of region specific forecasting models. The data-sets consist of 14-16 variables each, with the amounts of variables varying depending on the number of exchange connections to neighbouring zones.

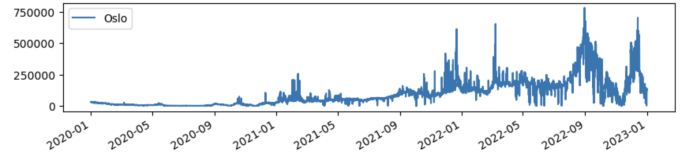


Figure 6. Historical Elspot prices for Oslo (NO1).

TABLE I. DESCRIPTION OF DATA (TARGET*).

Variable (units) [granularity]	Source
Elspot price (NOK/MWh) [h]	Nord Pool
Day-ahead Elspot price (NOK/MWh)[h]*	Nord Pool
Power production (MWh) [h]	Nord Pool
Power production prognosis (MWh) [h]	Nord Pool
Power exchange (MWh) [h]	Nord Pool
Power consumption (MWh) [h]	Nord Pool
Reservoir levels (GWh) [w]	Nord Pool
Reservoir capacity (GWh) [w]	Nord Pool
Gas price (NOK/mmbtu) [d]	Yahoo-finance
Oil price (NOK/barrel) [d]	Yahoo-finance
OSEBX price (NOK/OSEBX) [d]	Yahoo-finance
Air temperature (mean/degC) [d]	MET
Wind speed (mean/ms) [d]	MET
Precipitation (sum/mm) [d]	MET

Missing values occurred due to multiple reasons, such as changing time zones, observations at a lower frequency than the target values and stock exchanges being closed during weekends. Missing values due to these occurrences were appropriately imputed using interpolation, backward-fill or forward-fill. One example is the weather observation being recorded daily from hundreds of weather stations each day (see Figure 7), needing to be aggregated geographically to averages in each bidding-zone and filled for the 24-hours each day. Other preprocessing complexities include historical currency conversion of economic variables and handling large amounts of unstructured operational data.

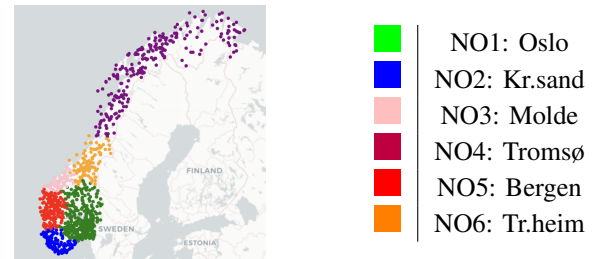


Figure 7. Locations of weather stations color-labeled by bidding zones.

The data is split into two sections, the first contains three years of data with 26 000+ price-observations and is allocated

for training and validation, the second is separated from the first and contains 4 months of recent and unseen data allocated for testing and evaluation. The date ranges are the following, 01.01.2020 00:00 - 29.12.2022 23:00 for train and validation, and 01.01.2023 00:00 - 30.03.2023 23:00 for the hold-out test set. Essentially, the train-test split contains the original time order and is not shuffled or re-ordered. Data is normalized using min-max scaling, this is done separately for the two sections in order to prevent introducing look-ahead-biases encoded in the scaling. In this research, all the data is scaled in order to help improve predictions and reduce risk of overfitting.

B. Data Analysis

To gain deeper insights into the relationships between independent and dependent variables, correlations and other descriptive statistics were computed. To further explore these relationships, Principal Component Analysis (PCA) was performed on the correlation matrices, projecting the variables onto a two-dimensional feature space, as illustrated in Figure 8. Singular Value Decomposition (SVD) provides a way to transform the data into a new coordinate system where the new axes (principal components) are linear combinations of the original variables, and the data can be represented in a lower-dimensional space with minimal loss of information. In all of the data-sets, the historical oil, gas and oseb prices are closely approximated in the feature-space. Weather variables, in particular precipitation is closely approximated to reservoir-levels in most bidding-zones and in some cases as in Oslo (Figure 8), precipitation and wind-speed is also relatively closely approximated to production.

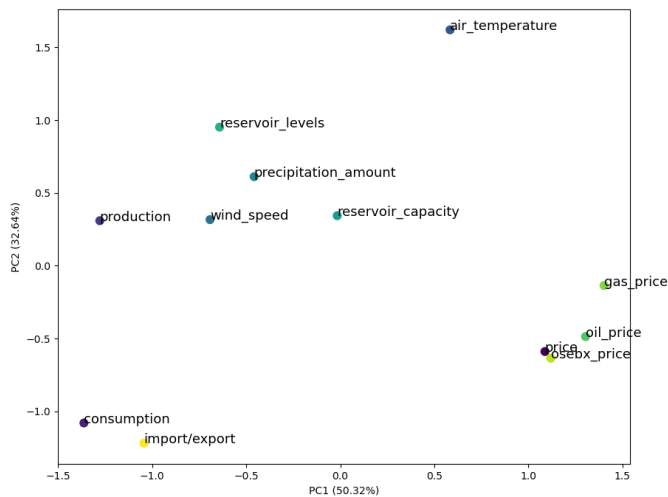


Figure 8. Approximation and projection of variables onto two-dimensional feature-space (PCA).

To better understand the price data we also conducted decomposition analysis and examined trends and seasonality, and potential memory effects within the time series by visualizing Auto-Correlation Functions (ACF) and Partial Auto-Correlation Functions (PACF). The stationarity of the data

was confirmed using Augmented Dickey-Fuller (ADF) statistics, ensuring the data is suitable for further modeling and predictive analysis. It is important to recognize that in real-world markets, the assumption of a stationary time series can be misleading. Financial and economic data, such as electricity prices or stock prices, are influenced by a range of external factors like policy changes, market shocks, and seasonal effects, which introduce non-stationary behavior over time.

C. Experiments

The experiments include a heuristic baseline and are compared against each other as opposed to previous experiments from related work. The persistence model is a naive approach that assumes the future price will be the same as the current price. In other words, it simply predicts the value of the target as the value of the target at time $t-24$. This model serves as a reference point for the performance of more complex models. The implementation of the persistence model is straightforward and can be easily achieved using any programming language or spreadsheet software. In this case, Python and the Pandas library was used to load and manipulate the data. The other models were trained and optimized in different ways due to their varying degree of complexity.

To fit an ARIMA model, optimal values for p , d , and q are typically determined by analyzing the data through various methods. This includes plotting the ACF and PACF, computing ADF statistics, and evaluating models using criteria like AIC or BIC. The optimal model is selected by fitting multiple models with different orders and choosing the one with the lowest AIC or BIC value.

$$AIC = -2 \ln(\hat{L}) + 2k \quad (6)$$

$$BIC = -2 \ln(\hat{L}) + k \ln(n) \quad (7)$$

Optimal parameters found:

$$p = 2, \quad d = 1, \quad q = 2$$

During training, XGBoost minimizes the loss function root mean squared error (RMSE), to improve prediction accuracy. Gradient boosting is employed to iteratively adjust the model's parameters and reduce the loss. One of the main advantages of XGBoost in this context is its ability to provide feature importance scores, which are derived from the model's decision trees. These scores offer interpretability in a TSF framework, allowing for an understanding of how different factors influence electricity prices over time. The feature importance scores were computed using the "gain" metric, it measures the contribution of each feature to reducing the loss function in the model. By analyzing these scores, it was possible to identify the most influential variables across different regions. In forecasting tasks, knowing the features have the most impact allows for targeted adjustments and improvements in the model. It also aids in validating and explaining model

predictions, enhancing transparency and trust in the model's outputs.

$$\mathcal{L}(\Theta) = \sqrt{\frac{1}{T} \sum_{t=1}^T \text{MSE}(y_t, \hat{y}_t)} \quad (8)$$

L2 Regularization:

$$\Omega(f) = \frac{\lambda}{2} \sum_{j=1}^T w_j^2 \quad (9)$$

where λ is the regularization strength.

Training neural networks involves feeding the model full cycles of training data, with each cycle called an epoch. After each epoch, the model updates its weights through a backward pass and optimization process to minimize the mean squared error (MSE) loss function. Configuring neural networks is complex due to the lack of a universal approach; instead, it requires systematic exploration of different configurations. This involves both dynamical exploration, assessing how the network behaves during training, and objective exploration, evaluating performance on validation or test sets. Finding the optimal number of epochs, number of hidden layers and number of neurons in each hidden layer must be explored. To address overfitting in neural networks, dropout regularization is applied to randomly drops weights and prevent excessive co-adaptation. Hyperparameter tuning, performed using the Optuna library, helps find the optimal configuration for aspects such as the number and shape of hidden layers, dropout rate, learning rate, batch size, and sequence length.

$$\mathcal{L}(\Theta) = \frac{1}{T} \sum_{t=1}^T \text{MSE}(y_t, \hat{y}_t) \quad (10)$$

Optimizer: Adam

First, the models are validated in the task of predicting the day-ahead hourly elspot prices on the validation set using a rolling forecast cross-validation (RFCV) scheme presented in Table II. These experiments provide information about the models' performance on a full year of daily-predictions with daily re-training. During validation, the error of the models is measured using RMSE. The errors are averaged by time of day; mornings (hours 6-12), mid-days (hours 12-15), evenings (hours 15-21) and nights (hours 21-6). An example of results from rolling forecasts origin validation with visualization from a sample period of 1 week including bar charts of aggregated time-of-day scores from the entire year are presented in Figure 9 (baseline results of aggregated RMSE are marked with red dashed lines for comparisons).

TABLE II. RFCV SCHEME (YYYY-MM-dd hh).

Fold	Train Start	Train End	Val Start	Val End
1	2020-01-01 00	2021-12-31 23	2022-01-01 00	2022-01-01 23
2	2020-01-01 00	2022-01-01 23	2022-01-02 00	2022-01-02 23
3	2020-01-01 00	2022-01-02 23	2022-01-03 00	2022-01-03 23
...
365	2020-01-01 00	2022-12-28 23	2022-12-29 00	2022-12-29 23

After validating the models on the last year of the train-set, they are then evaluated in their ability to extrapolate 24 time-steps ahead from the known spot-price during a 4-month out-of-sample period on a recent hold-out test-set from all the bidding-zones, with their weights and hyperparameters determined from training and tuning on the previous 3 years of data. The results of these experiments are presented in Table III, allowing for comprehensive analysis and review of the different modelling approaches in relation to the bidding zones and the addition of exogenous variables. The evaluation scheme of model performance consists of four different error terms; Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Average Percentage Error (MAPE) and Residual Sum of Squares (RSS). To gain a comprehensive understanding of the models' capacity for generalization and their ability to navigate the bias-variance trade-off, we seek to offer diverse viewpoints on the models' performance.

D. Results

This section outlines the key findings from the research project, incorporating results from our conference paper [1], which summarizes the performance of various predictive models. The scope has been extended to address the econometric ARIMA model, highlighting the importance of different features to facilitate better interpretation of model outputs. Detailed discussion and interpretation of the results follow in subsequent sections.

The actual vs. predicted NO₂ values for Kristiansand are presented for the time interval December 18th to December 24th, 2022. The top subplot (Figure 9a) shows the comparison between the actual and predicted values for a one-week period. The lower subplot (Figure 9b) aggregates the root mean square error (RMSE) by time of day for the entire year of 2022. Predictions from validation seem to be more accurate during mornings (6-12) and middays (12-15) as illustrated by the RMSE scores in Figure 9. However, none of the models consistently outperform the heuristic baseline across bidding zones and time-of-day during these experiments. The performance of different forecasting models (Heuristic, ARIMA, XGBoost, and LSTM) during out-of-sample evaluation is summarized in Table III, which reports the RMSE, MAE, MAPE, and RSS for each model with and without exogenous variables. The results cover the out-of-sample evaluation period from January 1st, 2023, to March 30th, 2023. The LSTM model in its multivariate configuration outperforms the other models for all aspects of error on the data-sets for bidding-zone NO₂ and NO₃. Surprisingly, the univariate LSTM outperforms the other models in all aspects of error for the bidding-zone NO₄. The final model to outperform the baseline for all aspects of error is the multivariate XGBoost model for the bidding-zone NO₆. For the remaining bidding-zones NO₁ and NO₅ there is no clear contender for best model performance. Among the models evaluated, LSTM and multivariate XGBoost models demonstrated superior performance, outperforming the baseline across all forecast criteria. These models successfully balanced the bias-variance trade-off, effectively capturing the

TABLE III. MODEL PERFORMANCE SUMMARY ON TEST SETS (01.01.2023 00:00 - 30.03.2023 23:00).

	Model	RMSE		MAE		MAPE		RSS	
		endog	w/ exog	endog	w/ exog	endog	w/ exog	endog	w/ exog
NO1	Heuristic	29469	/	19946	/	25.19%	/	18.7e ¹¹	/
	ARIMA	29469	44124	19946	36332	25.18%	36.84%	18.7e ¹¹	41.1e ¹¹
	XGBoost	29156	27052	20268	18838	26.22%	26.92%	17.9e ¹¹	15.5e¹¹
	LSTM	29174	21109	21035	20134	29.21%	26.69%	17.9e ¹¹	17.9e ¹¹
NO2	Heuristic	29474	/	19943	/	25.19%	/	18.7e ¹¹	/
	ARIMA	29474	37216	19943	27932	25.18%	30.86%	18.7e ¹¹	29.2e ¹¹
	XGBoost	29545	27259	20715	19266	26.77%	26.19%	18.4e ¹¹	15.7e ¹¹
	LSTM	28354	26431	20317	18173	29.09%	24.81%	16.9e ¹¹	14.9e¹¹
NO3	Heuristic	30448	/	21069	/	37.58%	/	20.0e ¹¹	/
	ARIMA	30448	44907	21069	34139	37.57%	47.61%	20.0e ¹¹	42.6e ¹¹
	XGBoost	28469	29069	19687	19666	35.79%	31.79%	17.1e ¹¹	17.8e ¹¹
	LSTM	28438	28381	20462	19228	40.91%	31.29%	17.1e ¹¹	16.6e¹¹
NO4	Heuristic	21456	/	11705	/	25.28%	/	99.4e ¹⁰	/
	ARIMA	21456	30875	11705	22584	25.27%	45.46%	99.4e ¹⁰	18.7e ¹⁶
	XGBoost	20592	23149	11424	12507	25.43%	29.35%	89.6e ¹⁰	11.3e ¹¹
	LSTM	19448	21675	10519	13155	22.76%	28.05%	79.9e¹⁰	96.1e ¹⁰
NO5	Heuristic	25240	/	16953	/	15.75%	/	13.7e¹⁰	/
	ARIMA	25240	27679	16953	19177	15.75%	17.28%	13.7e¹⁰	16.1e ¹¹
	XGBoost	24950	24156	17137	17018	16.27%	15.94%	13.1e ¹¹	12.3e ¹¹
	LSTM	25427	24584	18391	18189	17.80%	17.49%	13.6e ¹¹	12.9e ¹¹
NO6	Heuristic	30448	/	21069	/	37.58%	/	20.0e ¹¹	/
	ARIMA	30448	45333	21069	34465	37.57%	48.04%	20.0e ¹¹	43.4e ¹¹
	XGBoost	28469	28326	19687	19532	35.79%	31.70%	17.1e ¹¹	16.9e¹¹
	LSTM	28438	30100	20462	22870	40.91%	48.58%	17.1e ¹¹	19.1e ¹¹

complex data dynamics of EPF. Conversely, ARIMA models, while strong in interpretability and simplicity, faced limitations in out-of-sample extrapolation. The ARIMA models we configured struggled with longer inference horizons, exhibiting underfitting with endogenous variables alone and overfitting when incorporating exogenous variables. This performance discrepancy underscores the challenges of using ARIMA models for time series with complex and long-term dependencies, as their reliance on lagged values may fail to adequately capture and project complex price patterns. Table IV displays the feature importance scores for the XGBoost model across different regions. Key features such as energy prices, weather conditions, and exchange variables are ranked based on their contribution to the model's predictions. The importance of these features varies across the regions, with price and oil price generally being the most significant predictors.

E. Discussion

The observed regional differences between southern (NO1, NO2, NO5) and northern (NO3, NO4, NO6) bidding zones in Norway reveal important insights into the dynamics of the electricity markets. The southern zones' strong correlation with economic factors, particularly macroeconomic variables such as oil prices and global energy markets, indicates a heightened sensitivity to external shocks. This could explain the increased volatility in electricity prices in these regions, as they are more exposed to fluctuations in global supply and demand for energy commodities. The XGBoost feature importance analysis in Table IV supports this by highlighting the prominence of oil prices and other market-driven factors in the price forecasts for southern zones like Oslo (NO1).

These results suggest that economic policies and global market developments could have a disproportionate impact on electricity prices in the southern regions. In contrast, the northern zones show a more stable price formation process, driven by operational factors such as hydropower availability and local consumption patterns. This suggests that, despite the geographic proximity of the regions, the drivers of electricity prices differ significantly. The northern zones, less exposed to macroeconomic volatility, may experience more predictable and stable price trends, driven by supply-side considerations like hydropower generation and reservoir levels. This aligns with the relative stability observed in the northern zones, where local operational factors play a more substantial role. These regional distinctions underscore the importance of adopting tailored forecasting approaches for different bidding zones. For instance, models forecasting prices in southern regions could benefit from incorporating global economic indicators and commodity market trends, while models for northern regions should focus more on hydrological conditions and localized operational factors. Furthermore, the differing sensitivity of regions to price drivers has implications for policy-making, as energy regulation or economic policies that affect market dynamics may need to be region-specific to ensure stability and predictability in electricity prices across Norway. This analysis also raises questions about the resilience of the Norwegian electricity market to global economic shifts.

Lack of improvements over the baseline during validation seen in Figure 6 could be attributed to the disruptive prices in 2022, making it difficult for the models to fit the data comprehensively. Results from the out-of-sample evaluation

TABLE IV. FEATURE IMPORTANCE (GAIN) SCORES FOR XGBoost IN DIFFERENT REGIONS.

Feature	NO1	NO2	NO3	NO4	NO5	NO6
Price	33.05	41.01	8.04	4.54	33.97	8.6
Oil Price	12.2	4.61	3.73	0.51	3.76	3.26
Osebx Price	3.71	3.52	1.28	0.47	2.91	1.37
Gas Price	2.61	2.4	1.51	0.65	2.01	1.63
Reservoir Levels	1.94	2.77	1.35	0.68	1.21	1.05
Wind Speed	0.83	0.69	0.83	1.7	1.09	0.5
Air Temperature	0.45	1.24	0.77	1.23	0.59	2.1
Production	0.33	0.85	0.43	0.35	0.7	0.55
Production Forecast	0.51	1.24	0.91	1.03	0.66	0.62
Precipitation Amount	0.78	0.61	0.77	0.62	0.52	0.77
Consumption	0.36	0.23	0.92	0.46	0.32	0.43
Exchange NO1-NO2	0.24	0.26	-	-	-	-
Exchange NO1-NO3	0.11	-	0.28	-	-	-
Exchange NO1-NO5	0.29	-	-	-	0.24	-
Exchange NO1-NO6	-	-	-	-	-	0.25
Exchange NO1-SE3	0.23	-	-	-	-	-
Exchange NO2-NL	-	0.61	-	-	-	-
Exchange NO2-NO5	-	0.36	-	-	0.29	-
Exchange NO3-NO4	-	-	0.63	0.86	-	-
Exchange NO3-NO5	-	-	0.13	-	0.35	-
Exchange NO3-SE2	-	-	0.59	-	-	-
Exchange NO4-SE2	-	-	-	0.75	-	-
Exchange NO4-SE1	-	-	-	0.42	-	-
Exchange NO6-NO4	-	-	-	-	-	0.54
Exchange NO6-NO5	-	-	-	-	-	0.63
Exchange NO6-SE2	-	-	-	-	-	0.25

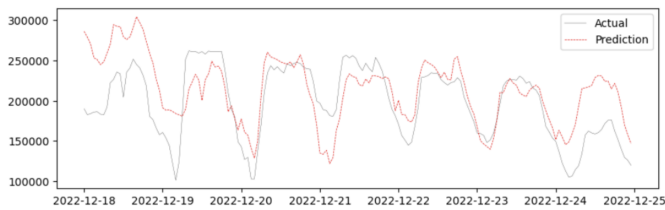
exhibit more promising improvements over the baseline. As seen in Table III, the LSTM and XGBoost models outperform the baseline across all evaluation criteria for most of the bidding-zones, meaning that they are able to balance between capturing price nuances while maintaining robustness to outliers. These results ultimately emphasize the potential of DL and ensemble ML techniques for capturing the complexities of EPF. The model performance across different bidding zones shows a mixed picture. The naive baseline model often performs well in terms of MAPE and RSS, suggesting it may be a strong benchmark for some zones. XGBoost generally excels in RMSE, indicating robust prediction accuracy, while also showing improvements with exogenous variables. The LSTM models, though slower to train and complex, tend to offer competitive performance, particularly in terms of RMSE and generalization, especially when configured with multiple variables. ARIMA shows variable results, sometimes overfitting or underfitting depending on the configuration. Regarding simplicity the ARIMA models stand out as contenders, often surpassing the baseline in validation. They offer a straightforward approach to TSF and ease of interpretation. However, challenges arise when extending these models to further out-of-sample inference. The interpretability of tree-based models like XGBoost provides significant advantages, particularly in understanding complex, non-linear relationships within the data. Unlike many black-box models, XGBoost offers clear insights into the features that are most influential in driving predictions. This interpretability is crucial for stakeholders who need to make informed decisions based on the

model's findings and ensures that the model's behavior aligns with domain knowledge and expectations. The variability in results across the different data-sets highlights the presence of unique characteristics for the distinct bidding-zones, with varying predictability, model-performances and optimal model-configurations. Highlighting the need for region-specific price-modelling and improving model generalization.

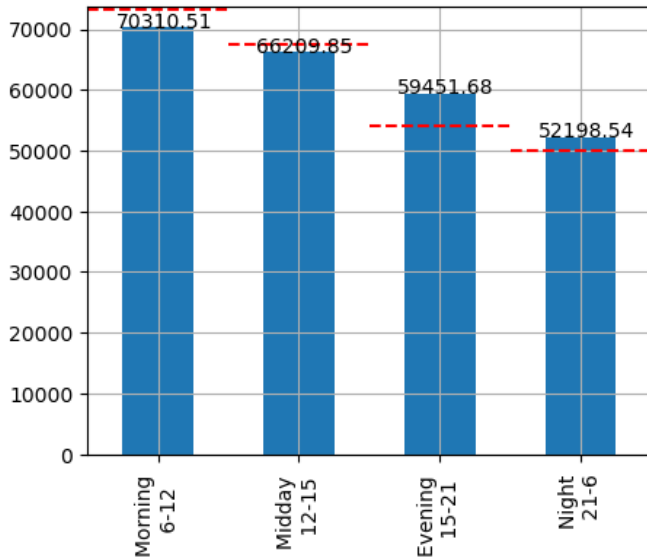
V. CONCLUSION AND FUTURE WORK

Forecasting day-ahead electricity prices plays a pivotal role in strategizing and balancing the supply and demand for the subsequent day, making it an essential area to delve into. In this paper, we introduce a framework to assess forecasting techniques across all Elspot markets in Norway, intimidating heuristic methods with more complex ARIMA models, advanced XGBoost and LSTM deep learning networks. Various models, including XGBoost and LSTM, show varying effectiveness across different bidding zones. XGBoost's interpretability aids in understanding non-linear relationships, while LSTM models demonstrate strong predictive capabilities they offer little insight into the patterns it learns from the data. Overall, the research underscores the importance of combining detailed analysis of price drivers with sophisticated modeling techniques to enhance the understanding and prediction of electricity markets.

The study's focus on the Norwegian market limits the generalizability of the findings. Future work should explore the applicability of the developed models to other electricity markets to assess their robustness across different contexts. Computational constraints and the omission of extensive



(a) Actual vs. prediction (18.12.2022 00:00 - 24.12.2022 23:00).



(b) Aggregated RMSE (01.01.2022 00:00 - 29.12.2022 23:00).

Figure 9. Rolling Forecast Origin Cross-validation of multivariate LSTM for Kristiansand (NO2).

feature engineering also highlight areas for improvement. Incorporating additional data sources and exploring hybrid models that combine various forecasting approaches could further refine prediction accuracy. The variability in results across the different data-sets highlights the presence of unique characteristics for the distinct bidding-zones, therefore, model generalization will be the focal point of our future research endeavors. In conclusion, this research contributes valuable insights into the Norwegian electricity markets and forecasting methodologies. Addressing the identified limitations and exploring future research directions will enhance the development of more accurate and reliable forecasting models, benefiting both researchers and practitioners in the field.

REFERENCES

- [1] M. W. Jensen, H. Ren, and A. Shalaginov, "Day-ahead Electricity Price Forecasting of Elspot Markets in Norway", *ENERGY 2024, The Fourteenth International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies*, vol. 14, pp. 1–6, 2024.
- [2] S.K Aggarwal, L. M Saini and Ashwani Kumar, "Electricity price forecasting in deregulated markets: A review and evaluation", *International Journal of Electrical Power & Energy Systems*, vol. 31, pp. 13-22, 2009.
- [3] N. Amjady and M. Hemmati, "Energy price forecasting - problems and proposals for such predictions", *IEEE Power and Energy Magazine*, vol. 4, pp. 20–29, 2006.
- [4] R. Beigaitė, T. Krilavicius and K. L. Man, "Electricity Price Forecasting for Nord Pool data", *International Conference on Platform Technology and Service (PlatCon)*, pp. 1–6, 2018.
- [5] K. Bitirgen and Basaran Filik, "Electricity price forecasting based on XGBoost and ARIMA algorithms", *BSEU Journal of engineering research and technology*, vol. 1, pp. 7–13, December 2020.
- [6] L. Breiman, *Classification and regression trees*, Routledge, 2017.
- [7] M. Castelli, A. Groznik, and A. Popovic, "Forecasting electricity prices: A machine learning approach", *Algorithms*, vol. 13, pp. 119, May 2020.
- [8] Z. Chang, Y. Zhang and W. Chen, "Electricity Price Prediction based on Hybrid Model of Adam optimized LSTM neural network and Wavelet Transform", *Energy*, pp. 187, 2019.
- [9] 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*, pp. 785–794, August 2016.
- [10] Y.-k. Chen, A. Hexeberg, K. E. Rosendahl and T. F. Bolkesjø, "Long-term trends of nordic power market: A review", *WIREs Energy and Environment*, vol. 10, pp. 413, 2021.
- [11] A. Ciarreta, M. P. Espinosa and C. Pizarro-Irizar, "Is green energy expensive? empirical evidence from the Spanish electricity market", *Energy Policy*, vol. 69, pp. 205–215, 2014.
- [12] A. Conejo, M. Plazas, R. Espinola and A. Molina, "Day-ahead electricity price forecasting using the wavelet transform and ARIMA models", *IEEE Transactions on Power Systems*, vol. 20, pp. 1035–1042, 2005.
- [13] J. Contreras, R. Espinola, F. Nogales and A. Conejo, "Arima models to predict next-day electricity prices", *IEEE Transactions on Power Systems*, vol. 18, pp. 1014–1020, September 2003.
- [14] A. Creti and F. Fontini, *Economics of electricity*, Cambridge University Press, 2019.
- [15] A. Cruz, A. Munoz, J. L. Zamora and R. Espinola, "The effect of wind generation and weekday on spanish electricity spot price forecasting", *Electric Power Systems Research*, vol. 81, pp. 1924–1935, 2011.
- [16] C. Defeuilley, "Retail competition in electricity markets", *Energy Policy*, vol. 37, pp. 377–386, 2009.
- [17] G. Erdmann, "Economics of electricity", *EPJ Web of Conferences*, vol. 98:06001, January 2015.
- [18] G. Gao, K. Lo and F. Fan, "Comparison of ARIMA and ann models used in electricity price forecasting for power market", *Energy and Power Engineering*, vol. 09, pp. 120–126, January 2017.
- [19] P. S. Georgilakis, "Market clearing price forecasting in deregulated electricity markets using adaptively trained neural networks", *SETN: Hellenic Conference on Artificial Intelligence*, vol. 3955, pp. 56–66, 2006.
- [20] L. Grossi and F. Nan, "Robust forecasting of electricity prices: Simulations, models and the impact of renewable sources", *Technological Forecasting and Social Change*, vol. 141, pp. 305–318, 2019.
- [21] J.-J. Guo and P. Luh, "Selecting input factors for clusters of Gaussian radial basis function networks to improve market clearing price prediction", *IEEE Transactions on Power Systems*, vol. 18, pp. 665–672, June 2003.
- [22] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural computation*, vol. 9, pp. 1735–1780, 1997.
- [23] Z. Hu, L. Yang, Z. Wang, D. Gan, W. Sun and K. Wang, "A game-theoretic model for electricity markets with tight capacity constraints", *International Journal of Electrical Power & Energy Systems*, vol. 30, pp. 207–215, 2008.
- [24] J. Lago, G. Marcjasz, B. De Schutter, and R. Weron, "Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark", *Applied Energy*, vol. 293, pp. 116983, 2021.
- [25] C. Li and S. Wang, "Next-day power market clearing price forecasting using artificial fish-swarm based neural network", *ISNN: Advances in Neural Networks*, pp. 1290–1295, 2006.
- [26] P. Mandal, T. Senjyu and T. Funabashi, "Neural networks approach to forecast several hour ahead electricity prices and loads in deregulated market", *Energy Conversion and Management*, vol. 47, pp. 2128–2142, 2006.
- [27] M. S. Nazar, A. E. Fard, A. Heidari, M. Shafie-khah and J. P. Catalao, "Hybrid model using three-stage algorithm for simultaneous load and price forecasting", *Electric Power Systems Research*, vol. 165, pp. 214–228, 2018.
- [28] E. Raviv, K. E. Bouwman and D. Van Dijk, "Forecasting day-ahead electricity prices: Utilizing hourly prices", *Energy Economics*, vol. 50, pp. 227–239, 2015.

- [29] K. Skytte, "The regulating power market on the nordic power exchange nord pool: an econometric analysis", *Energy Economics*, vol. 21, pp. 295–308, 1999.
- [30] B. A. Souhir, B. Heni and B. Lotfi, "Price risk and hedging strategies in nord pool electricity market evidence with sector indexes", *Energy Economics*, vol. 80, pp. 635–655, 2019.
- [31] R. Weron, *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach*, John Wiley Sons, 2006.
- [32] R. Weron and A. Misiorek, "Forecasting spot electricity prices with time series models", *Proceedings of the European Electricity Market EEM-05 Conference*, pp. 133–141, May 2005.
- [33] R. Weron and M. Zator, "Revisiting the relationship between spot and futures prices in the nord pool electricity market", *Energy Economics*, vol. 44, pp. 178–190, 2014.
- [34] Wikipedia, "Time series," *Wikipedia, The Free Encyclopedia*, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Time_series&oldid=1193976963. [Accessed: 2024-02-09].
- [35] Z. Yang, L. Ce, and L. Lian, "Electricity price forecasting by a hybrid model, combining wavelet transform, arma and kernel-based extreme learning machine methods", *Applied Energy*, vol. 190, pp. 291–305, 2017.

Improving Effectiveness and Performance Based on Dimensionality Reduction of CCD Image Features in Fall Armyworm's Control

Alex B. Bertolla^{1,2} and Paulo E. Cruvinel^{1,2}

¹Embrapa Instrumentation, São Carlos, SP, Brazil

²Federal University of São Carlos - Post Graduation Program in Computer Science, São Carlos, SP, Brazil

E-mails: alex.bertolla@embrapa.br, paulo.cruvinel@embrapa.br

Abstract—The pest control in agriculture based on digital imaging sensors has increased significantly in the past decades. Such a strategy has become possible due to the continuous improvements in computational intelligence and machine learning techniques. However, the demand for analyzing and processing such an amount of data generated by these sensors has become a challenge due to the high dimensionality. This article presents a study on the dimensionality reduction of features from digital images acquired with a Charge-Coupled Devices sensor in an agricultural field, to choose the optimal number of principal components for reducing feature dimensionality. It also presents a machine-learning method for the pattern recognition of this species of caterpillar (Fall armyworms - *Spodoptera frugiperda*) in its different growth stages. In such a context, selecting the optimal number of principal components for dimensionality reduction, retaining only the necessary information associated with the main variables that describe the object of interest. The results have shown that using Hu invariant moments for feature extraction, dimensionality reduction was possible for all analyzed cases, leading to 80% of the original data. In this context, it was possible to preserve the semantic characteristics collected by the sensor. Support Vector Machine classifiers have reached more than 70% of accuracy and more than 80% of precision. Moreover, the performance of the classifiers was 30% faster when working with the dimensionality reduced of the feature vector than when working with the original data.

Keywords—ccd sensor; digital image; feature extraction; dimensionality reduction; principal component analysis.

I. INTRODUCTION

In agriculture pest control plays an important role. In maize production, the Fall armyworm (*Spodoptera frugiperda*) has been requiring special attention, since it sponsors significant losses in production. In such a context, a previous study has been presented at the Ninth International Conference on Advances in Sensors, Actuators, Metering and Sensing (ALLSENSORS 2024) [1].

Charge-coupled devices (CCD) are the most used imaging sensors for digital image acquisition. They have built-in frame capture systems and the analog-to-digital conversion is done in the sensor itself [2].

CCD's sensors have been used in such ways to acquire images for different purposes. In agriculture, those sensors are usually used to capture images of pests and diseases [3] [4].

Due to the complex and high dimensions of the data captured by those sensors, storing and processing the amount

of data acquired has become a challenging task [5], known as the curse of dimensionality [6]. This phenomenon is related to the fact, that with a certain degree of accuracy from a function estimation, the number of variables increases as the number of samples also has to increase [7].

To solve the issue of the curse of dimensionality, different methods based on dimensionality reduction techniques have been proposed [8]. These methods transform the original high-dimensional data into a new reduced dataset, removing the redundant and non-relevant features [9]. Dimensionality reduction algorithms allow an efficient reduction of the number of variables, and if applied before machine learning models can avoid overfitting.

In the literature, it is possible to find several available researches about dimensionality reduction techniques for different types of data, such as Principal Component Analysis (PCA) introduced in 1901 by Karl Pearson [10], and its variations [11], Linear Discriminant Analysis (LDA) [12], Singular Value Decomposition (SVD) [13] and Isometric Mapping (ISOMAP) [14], a non-linear dimensionality reduction method based on the spectral theory, which tries to preserve the geodesic distances in the lower dimension.

PCA is a linear dimension reduction technique and is the most predominant method applied [15], and was considered to compose this work.

This paper presents a method for dimensionality reduction optimization when using CCD sensor-based images to control Fall armyworms in agriculture. In fact, the task of image classification allows the machine to understand what type of information is contained in an image, on the other hand, semantic segmentation methods allow the precise location of different kinds of visual information, as well as each begins and ends. Besides, it also presents a case study based on selecting Support Vector Machine (SVM) classifiers to evaluate the reduced features and their relation with the original data.

After the introduction, the remainder of the paper is organized as follows: Section II describes the work methodology; Section III shows the results, Section IV the discussion of the experiments; and finally, Section IV presents the conclusion of this paper and suggestions for future works.

II. METHODS

All the experiments have been performed in Python, i.e., by using both the image processing and machine learning libraries in openCV, as well as scikit-image and scikit-learn algorithms respectively. We also have considered an operating platform with a 64-bit CPU Intel (R) model Core(TM) i7-970, 16Gb RAM, and operational system Microsoft Windows 11.

A. Digital Image Sensor and Dataset

A digital image can be defined as a bi-dimensional function $f(x, y)$, where (x, y) are the intensity positions, defined as pixel [16]. CCD's sensors can capture images in different color spaces, however, the most common color space is the Red Green, and Blue (RGB), representing the visible spectrum [17].

Table I presents the features of the images acquired using the CCD sensor.

TABLE I
IMAGE FEATURES ACQUIRED BY CCD SENSOR

Image type	JPG / JPEG
Color space	RGB
Width	3072 pixels
Height	2048 pixels
Resolution	72 pixels per inch (ppi)
Pixel size	0.35mm

Regarding the image acquisition, a dataset was generated using a CCD sensor. This dataset is composed of the Fall armyworm images in real maize crops, where the pest was found both in leaves and cobs maize.

B. Feature Extraction

The Hu invariant moments descriptor was considered for the extraction of the geometric features of the pest. For the calculation of the seven invariant moments of Hu, it is necessary, a priori, to calculate the two-dimensional moments, that is, the central moments and normalized central moments [18]. Two-dimensional moments are understood to be the polynomial functions projected onto a 2D image, $f(x, y)$, and size $M \times N$ and order $(p + q)$.

The normalized central moments allow the central moments to be invariant to scale transformations, being defined by:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^\gamma} \quad \text{for } \mu_{00} \neq 0 \quad (1)$$

where γ is defined as:

$$\gamma = \frac{p+q}{2} + 1 \quad (2)$$

for $p + q = 2, 3, \dots$, positive integers $\in \mathbb{Z}$.

In this way, the invariant moments can be calculated considering:

$$\phi_1 = \eta_{20} + \eta_{02} \quad (3)$$

$$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \quad (4)$$

$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \quad (5)$$

$$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \quad (6)$$

$$\phi_5 = \frac{(\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})}{[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})} [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \quad (7)$$

$$\phi_6 = \frac{(\eta_{20} - \eta_{02})}{[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})\eta_{21} + \eta_{03}} \quad (8)$$

$$\phi_7 = \frac{(3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})}{[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})} [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \quad (9)$$

Neither of the seven Hu invariant moments is directly related to the size of an object in an image. However, the size of an object can be indirectly inferred through either the first or fourth moment [19].

After the features are extracted using the methods considered, a single feature vector is organized. Then, to reduce its dimensionality, PCA is applied [20].

C. Principal Components Analysis

PCA considers an array \mathbf{X} of data with n samples representing the number of observations and m independent variables [21], that is:

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix} \quad (10)$$

Herein, the principal components are obtained for a set of m variables X_1, X_2, \dots, X_m with means $\mu_1, \mu_2, \dots, \mu_m$ and variance $\sigma_1^2, \sigma_2^2, \dots, \sigma_m^2$, which are independent and have covariance between the n -th and m -th variable [9], in the form:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & \cdots & \sigma_{1m}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{n1}^2 & \cdots & \sigma_{nm}^2 \end{bmatrix} \quad (11)$$

where $\mathbf{\Sigma}$ represents the covariance matrix. To do this, the pairs of eigenvalues and eigenvectors are found $(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_m, e_m)$, where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$ and associated with $\mathbf{\Sigma}$ [22], where the i -th principal component is defined by:

$$Z_i = e_{i1}X_1 + e_{i2}X_2 + \dots + e_{im}X_m \quad (12)$$

where Z_i is the i -th principal component. The objective is to maximize the variance of Z_i , as:

$$\text{Var}(Z_i) = \text{Var}(e_i' \mathbf{X}) = e_i' \text{Var}(\mathbf{X}) e_i = e_i' \mathbf{\Sigma} e_i \quad (13)$$

where $i = 1, \dots, m$. Thus, the spectral decomposition of the matrix Σ is given by $\Sigma = \mathbf{P}\mathbf{\Lambda}\mathbf{P}'$, where \mathbf{P} is the composite matrix by the eigenvectors of Σ , and $\mathbf{\Lambda}$ the diagonal matrix of eigenvalues of Σ [23]. Thus, it has to be:

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_m \end{bmatrix} \quad (14)$$

In general, the principal component of greatest importance is defined as the one with the greatest variance, which explains the maximum variability in the data vector. The second most important component is the component with the second highest variance, and so on, up to the least important component [13].

Likewise, the normalized eigenvectors represent the main components that constitute the feature vector with reduced dimension. Besides, such reduced components are used to describe the acquired images. Additionally, the reduced features are used for the recognition of the patterns of Fall armyworm (*Spodoptera frugiperda*), i.e., useful consideration for both cases, leaf or cob maizes.

D. Machine Learning and Pattern Recognition

The ability of a computational system to improve the performance of a task based on experience can be defined as machine learning (ML), performed through either supervised or unsupervised learning methods [24].

In such a context, the feature vector, with reduced dimensionality, was used such that the classification was considered according to its position in the feature space. Thus, groups composed of similar characteristics could be identified and classified using support vector machine (SVM) classifiers [25].

The SVM's classifier can be established based on linear behavior or even non-linear behavior. Classifiers with linear behavior use a hyperplane that maximizes the separation between two classes from a training dataset and their respective labels [26]. In this case, the hyperplane is defined by:

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \quad (15)$$

where \mathbf{w} is the normal vector to the hyperplane, $\mathbf{w} \cdot \mathbf{x}$ is the dot product of the vectors \mathbf{w} and \mathbf{x} , and b is a fit term. Thus, Equation (16) divides the input space \mathbf{X} into two regions, as follows:

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x}_i + b &\geq 1 & \text{se } y_i &= +1 \\ \mathbf{w} \cdot \mathbf{x}_i + b &\leq -1 & \text{se } y_i &= -1 \end{aligned} \quad (16)$$

which can be summarized as:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0, \quad \forall (\mathbf{x}_i, y_i) \in \mathbf{X} \quad (17)$$

Linearly separable datasets are classified efficiently by linear SVMs with some error tolerance with smooth margins. However, in several cases, it is not possible to efficiently classify training data using this modality of a hyperplane [26], requiring the use of interpolation functions that allow

the operation in larger space, that is, using non-linear SVM classifiers.

In such a manner, SVMS can deal with non-linear problems through a Φ function, mapping the dataset from its original space (input space) to a larger space (input space, characteristics) [27], characterizing a non-linear SVM classifier.

Besides, from the choice of Φ , the training data set \mathbf{x} , in its input space R^2 , is scaled to the feature space R^3 , as:

$$\Phi(\mathbf{x}) = \Phi(x_1, x_2) = (x^2, \sqrt{2}x_1x_2, x_2^2), \quad (18)$$

$$\begin{aligned} h(\mathbf{x}) &= \mathbf{w} \cdot \Phi(\mathbf{x}) + b = \\ w_1x_1^2 + w_2\sqrt{2}x_1x_2 + w_3x_2^2 + b &= 0 \end{aligned} \quad (19)$$

The data are initially mapped to a larger space, then a linear SVM is applied over the new space. A hyperplane is then found with a greater margin of separation, ensuring better generalization [28].

Thus, the classifier obtained becomes:

$$g(x) = \text{sgn}(h(x)) = \text{sgn}\left(\sum_{x_i \in SV} \alpha_i^* y_i \Phi(x_i) \cdot \Phi(x) + b^*\right) \quad (20)$$

where b^* is calculated as:

$$b^* = \frac{1}{n_{SV:\alpha^* < C} \sum_{x_j \in SV:\alpha_j^* < C} \left(\frac{1}{y_j} - \sum_{x_i \in SV} \alpha_i^* y_i \Phi(x_i) \cdot \Phi(x_j)\right)} \quad (21)$$

Given that the feature space can be in a very high dimension, the calculation of Φ might be extremely costly, or even unfeasible. However, the only necessary information about the mapping is the calculation of the scalar products between the data in the feature space, obtained through function kernels [26].

Table II presents the kernels analyzed to validate the developed method.

TABLE II
SUPPORT VECTOR MACHINE FUNCTION KERNELS

Kernel	Function $K(x_i, x_j)$	Parameters
Polynomial	$(\delta(\mathbf{x}_i \cdot \mathbf{x}_j) + \kappa)^d$	δ, κ, e, d
Radial basis function kernel	$\exp(-\sigma \ \mathbf{x}_i - \mathbf{x}_j\ ^2)$	δ, κ, e, d
Sigmoidal	$\tanh(\delta(\mathbf{x}_i \cdot \mathbf{x}_j) + \kappa)$	δ, e, κ

The Radial basis function kernel (RBF), which is based on a Gaussian function, has been chosen for the ML process, to classify the Fall armyworm growth stage.

Accuracy and precision metrics have been measured for validation of the SVM classifiers, illustrated by both the confusion matrix and the receiver operating characteristic (ROC) curve.

III. RESULTS

Figure 1(a to e) illustrates one example of each stage of growth, also named Instar, Figure 1(f) illustrates two different Instar in the same image.

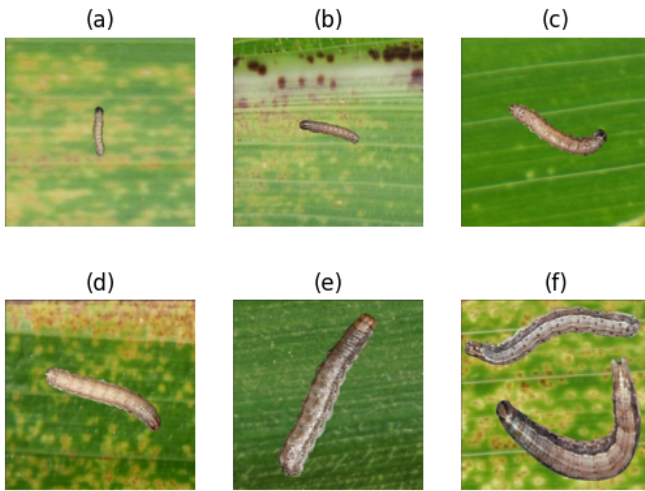


Figure 1. Fall armyworm (*Spodoptera frugiperda*) in different stages of growth.

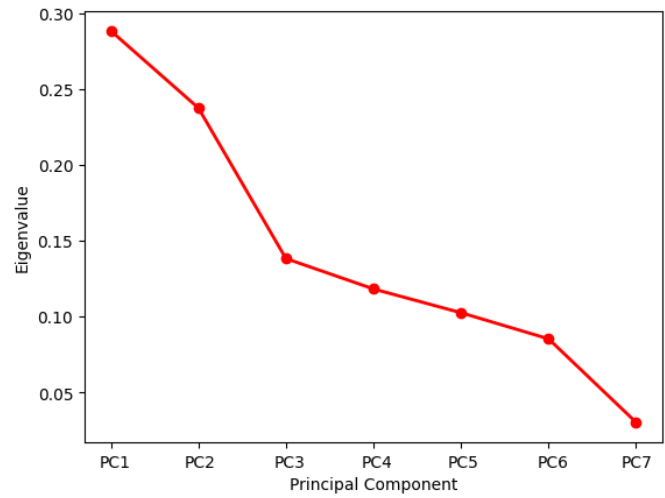


Figure 2. Scree plot.

TABLE III
FEATURE VECTOR COMPOSED OF HU INVARIANT MOMENTS. EXAMPLE OF THREE IMAGES

Hu invariant moments	Images		
	Image 1	Image 2	Image 3
ϕ_1	6.692	6.6178	6.524
ϕ_2	13.581	13.424	19.102
ϕ_3	24.321	23.944	22.370
ϕ_4	25.919	26.245	23.445
ϕ_5	51.517	-52.023	46.665
ϕ_6	34.307	-33.305	-34.728
ϕ_7	-51.458	-51.556	47.656

Table III presents the seven Hu invariant moments, as examples, from three different images, which were processed using the dataset.

Table IV presents the normalized seven Hu invariant moments from three different images.

TABLE IV
NORMALIZED FEATURE VECTOR. EXAMPLE OF THREE IMAGES

Hu invariant moments	Images		
	Image 1	Image 2	Image 3
ϕ_1	0.274	0.162	0.021
ϕ_2	-1.048	-1.121	1.496
ϕ_3	1.035	0.817	-0.092
ϕ_4	1.408	1.669	-0.575
ϕ_5	0.719	-1.607	0.610
ϕ_6	0.808	-1.289	-1.333
ϕ_7	-0.863	-0.865	1.199

Figure 2 shows the scree plot of the variance ratio.

Table V presents the maximum variance to each of the four principal components concerning the original data.

Figure 3 illustrates the maximum variance to each of the four principal components concerning the original data with the absolute values.

TABLE V
MAXIMUM VARIATION OF DATA IN RELATION TO EACH PRINCIPAL COMPONENT. BASED ON FOUR PRINCIPAL COMPONENTS.

Hu invariant moments	Principal components			
	PC 1	PC 2	PC 3	PC 4
ϕ_1	-0.147	-0.630	-0.120	-0.568
ϕ_2	0.501	-0.295	0.036	0.230
ϕ_3	-0.395	-0.424	0.087	0.027
ϕ_4	-0.378	0.501	-0.158	-0.479
ϕ_5	-0.376	-0.277	-0.310	0.286
ϕ_6	-0.355	-0.011	0.865	0.127
ϕ_7	0.398	-0.078	0.325	-0.543

Table VI presents the values of the four principal components.

TABLE VI
FEATURE VECTOR COMPOSED OF FOUR PRINCIPAL COMPONENTS. EXAMPLE OF THREE IMAGES

Principal components	Images		
	Image 1	Image 2	Image 3
PC1	0.333	2.121	-0.742
PC2	-2.280	-1.156	1.306
PC3	0.551	-1.243	-0.528
PC4	0.181	-1.193	0.012

The distribution of the variation of the four principal component values is illustrated in Figure 4.

This information can be observed in Figure 5, which illustrates a boxplot chart of the four principal component values and their distribution.

After the original data were reduced to two and four principal components, the experiments performed for classification considered SVM with a Gaussian function kernel, and the feature vector with reduced dimensionality was split into 70% for training and 30% for testing.

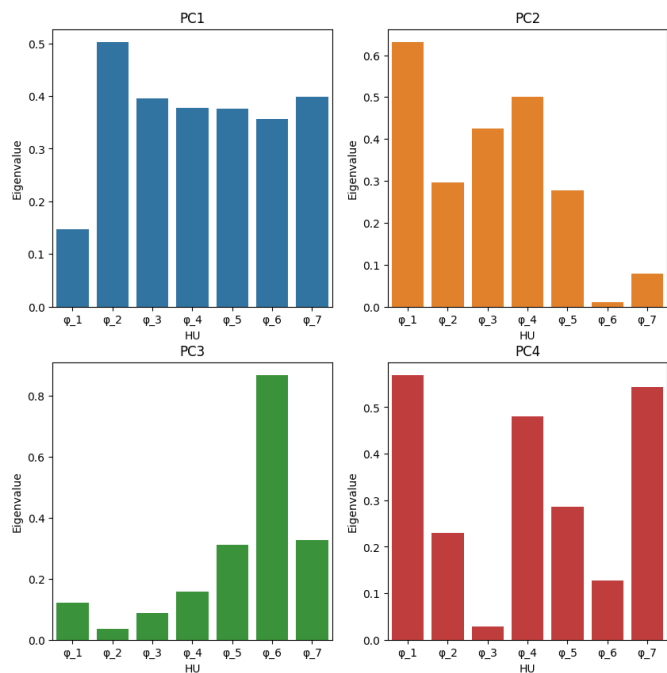


Figure 3. Maximum of data variation concerning each principal component, based on four principal components.

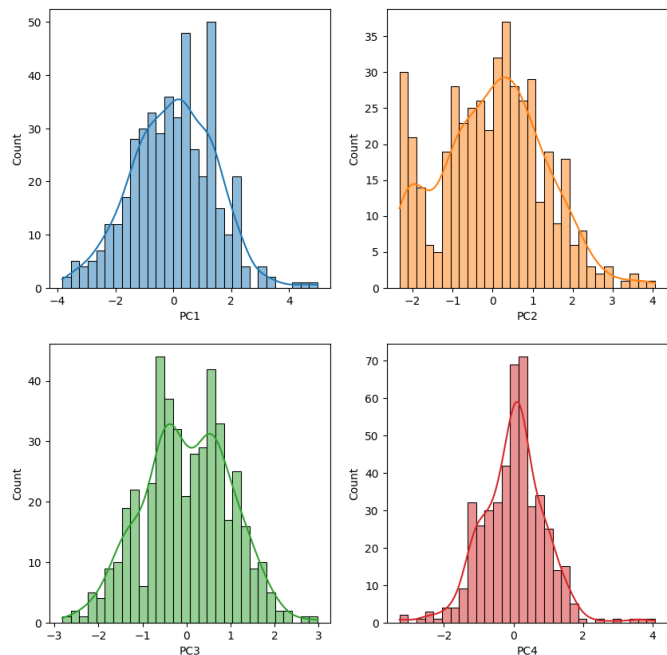


Figure 4. Histogram of distribution of the four principal components values.

Table VII presents the results of the classification of the five different stages of growth of the Fall armyworm based on four principal components.

Figure 6 illustrates the confusion matrix based on the data presented in Table VII.

Figure 7 illustrates the ROC curve based on the data presented in Table VII.

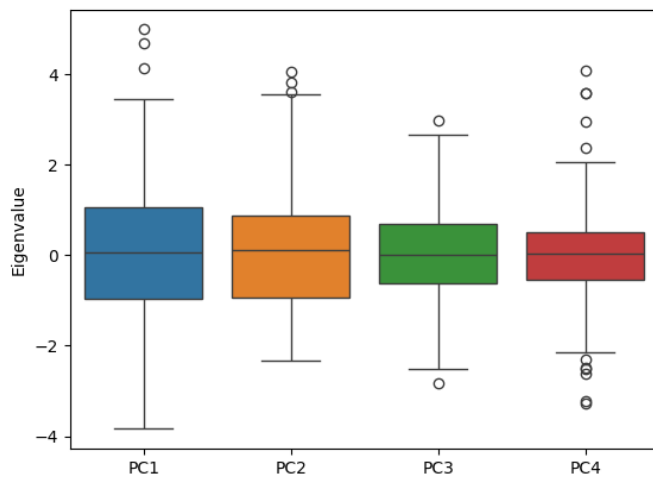


Figure 5. Boxplot of four principal components values.

TABLE VII
THE FALL ARMYWORM CLASSIFICATION RESULTS BASED ON FOUR PRINCIPAL COMPONENTS

Instar	Precision	Recall	F1-score
1	0.63	0.74	0.68
2	0.72	0.76	0.74
3	0.72	0.65	0.68
4	0.69	0.75	0.72
5	0.86	0.68	0.76

Based on the use of PCA the vector with the features for pattern recognition has been reduced in dimensionality, after that, the resultant vector was classified using an SVM classifier, i.e., having Gaussian kernel.

Table VIII presents the results of the classification of the five different stages of growth of the Fall armyworm.

TABLE VIII
THE FALL ARMYWORM CLASSIFICATION RESULTS BASED ON TWO PRINCIPAL COMPONENTS

Instar	Precision	Recall	F1-score
1	0.44	0.59	0.50
2	0.34	0.54	0.42
3	0.50	0.28	0.36
4	0.52	0.42	0.47
5	0.68	0.54	0.60

Figure 8 illustrates the confusion matrix based on the data presented in Table VIII.

Figure 9 illustrates the ROC curve based on the data presented in Table VIII.

The SVM classifier based on two principal components performed the classification of the five different stages of growth of the Fall armyworm with an accuracy of 47%.

Figure 10 illustrates the performance of the SVM classifier considering these selected components.

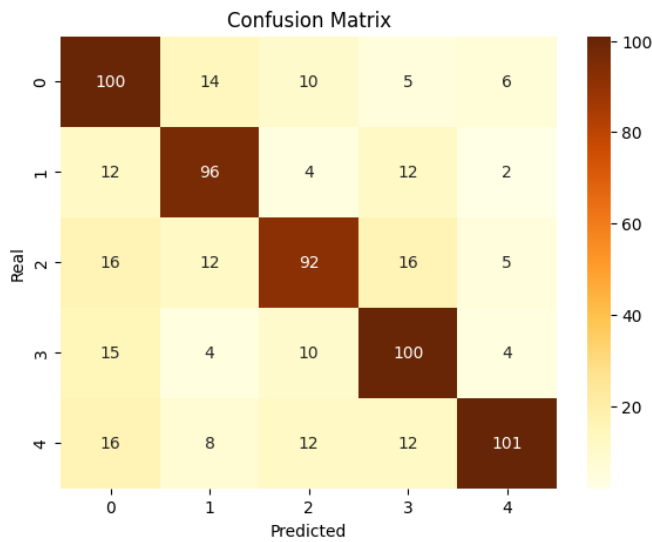


Figure 6. Confusion matrix for four principal components.

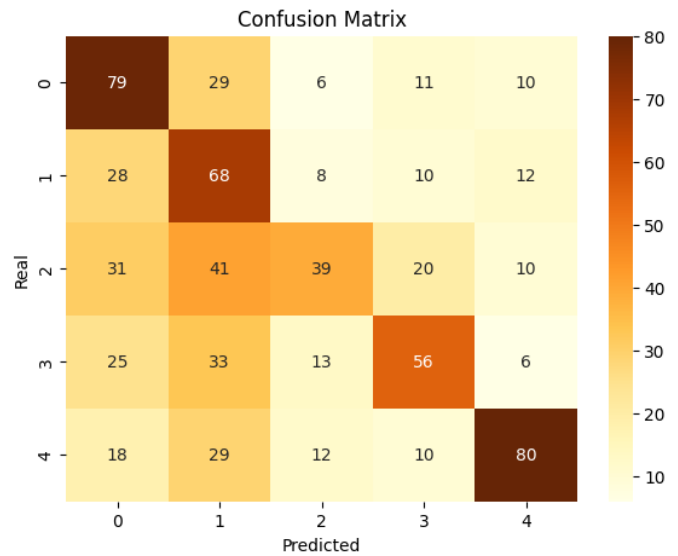


Figure 8. Confusion matrix for two principal components.

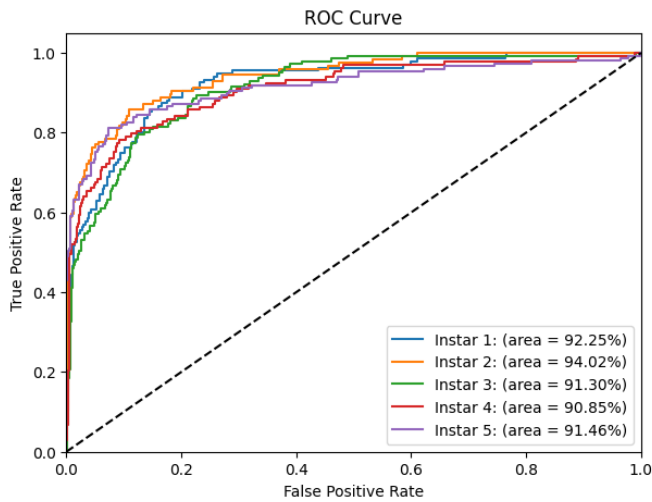


Figure 7. ROC curve for four principal components.

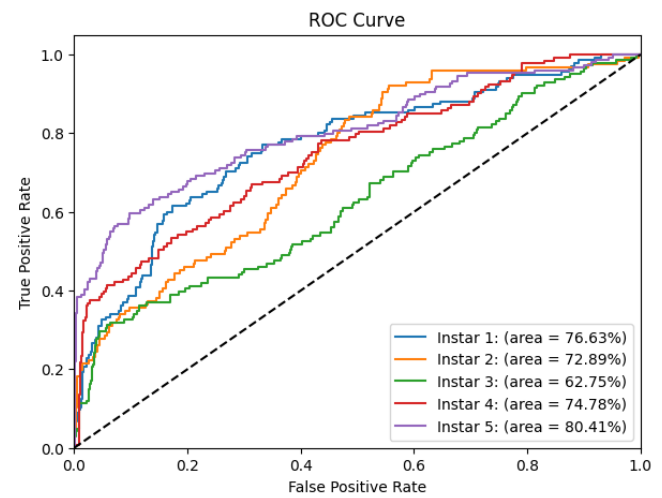


Figure 9. ROC curve for four principal components.

IV. DISCUSSION

For this study, an image dataset composed of 2280 images acquired with CCD's sensor was used. These images represent the Fall armyworm (*Spodoptera frugiperda*) acquired in a real environment of maize crop in its five different stages of growth, grouped in 456 images for each stage.

Considering all images contain at least one Fall armyworm in different stages, the Hu invariant moments descriptor has been considered for instance. Thus, for each image of the Fall armyworm, a feature vector was generated, containing the seven Hu invariant moments ($\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6$ and ϕ_7), which are related to the shape and geometrical features of this pest. The features contained in these vectors will allow the classification of the Fall armyworm (*Spodoptera frugiperda*) in its different stages of growth.

As the values of the feature vectors were in different scales, it was necessary to normalize them. To generate a database of characteristics of the Fall armyworm (*Spodoptera frugiperda*), the feature vectors referring to each image were saved on disk.

The removal of duplicated and less significant information, based on the application of PCA, has allowed improvements in computational performance. In fact, the appropriate number of principal components that explain the original data were considered.

Therefore, it is possible to infer that by applying two to four principal components it is possible to explain almost 55% to 80% of the variability of the original data. Considering that, the experiments were based on four principal components.

As already discussed, neither of the seven invariant moments is directly related to the size of an object. However, the first and the fourth moments can be used to infer the size of

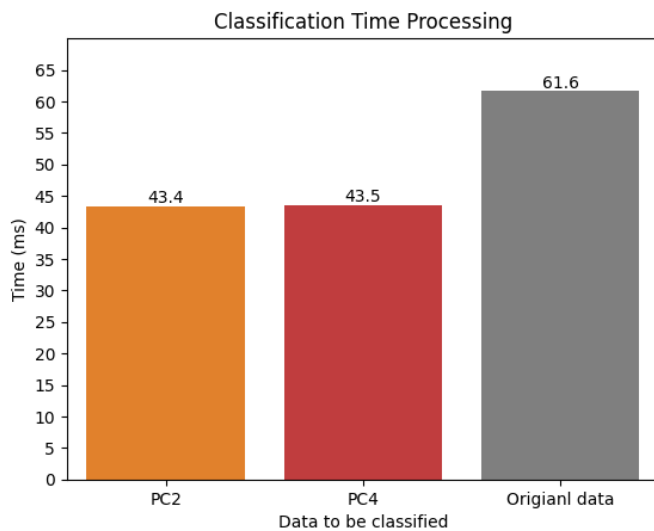


Figure 10. Classification time processing considering two and four principal components and the original data.

an object in an image.

Moreover, through the maximum variation ratio metric it is possible to measure the weight of each of Hu invariant moments in each principal component.

Even though the maximum variation ratio presents some negative values, the weights for each principal component are considered absolute values. For example, in PC2, the first moment (ϕ_1) has the highest weight.

The experiment with four principal components showed that, as can be visualized in Figure 3, to have the most representative weights either from the first moment or the fourth moment, it was necessary to work with two or four principal components.

The feature vector's dimensionality reduction was performed Based on prior experiments. Table VI presents the values of the four principal components.

Once the values of the four principal components are obtained, it is necessary to evaluate if it is sufficient to work with two principal components, or whether it should be considered four principal components. For this purpose, the maximum variation of each principal component should be considered concerning the original data, the extent to which the principal components could explain the original data and the minimum error.

This experiment demonstrated that working with four principal components was the ideal option. Because both the first and fourth moments are very representative, four principal components can explain 80% of the original data, and even with a low increase in the error, it is not considerable to decrease the estimation.

Furthermore, SVM classifiers with Gaussian function kernel have been considered for machine learning processes. Both accuracy and precision were taken into account to validate the classification of the Fall armyworm, with features vector dimensionality reduced.

The features vector composed of two and four principal components was split into two segments, one for training the SVM classifier with the proportion of 70% and the other one, with 30% for testing purposes.

Results obtained in the classification process with two principal components and illustrated in the ROC curve presented in Figure 9, showed the SVM classifier has performed satisfactorily, with the accuracy rate of 30% demonstrates that working with two principal components might have satisfactory performance for representing the original data, however, in the proposed scenario, the classification of the different stage of growth of the Fall armyworm with two principal components might not be enough.

On the other hand, when the classification process was performed with four principal components, the ROC curve illustrated in Figure 7 showed the classification of the five different stages the accuracy ratio assessed increased to 71%, obtained an efficient classification of patterns of the Fall armyworm.

Finally, Figure 10 shows the performance of the SVM classifiers in milliseconds to execute the testing of the dataset classification. As expected, the original feature vector took more time to be processed, 61.1 milliseconds, while the feature vector with reduced dimensionality took less time to be classified, 43.5 and 43.4 for PC4 and PC2, respectively.

V. CONCLUSION AND FUTURE WORK

This paper presented a study of dimensionality reduction using Principal Components Analysis (PCA), considering feature vectors composed of extracted Hu invariant moments.

Before measuring the number of principal components necessary to represent the original data from the Fall armyworm digital images, the feature vectors were normalized, to obtain all the seven Hu invariant moments.

The measure of the explained variance ratio to the original data was applied to verify the quantity number of principal components necessary to explain the maximum of the original data.

In addition, the first and fourth invariant moments were used to infer the estimated size of the Fall armyworm (*Spodoptera frugiperda*) in the images.

Likewise, the measure of the maximum variation of each principal component, concerning each Hu invariant moment, was performed to find how much these moments contribute to recognizing the main features acquired with the CCD's sensor.

The measurements have shown that computing two to four principal components was sufficient to explain 55% to 80% of the original data, and either the first or fourth moments were contained in two and four principal components.

Despite seven invariant moments being used, such analysis led to the conclusion that when using 4 principal components, one may achieve the explanation of 80% for the original data, with low error, as well as, not a significative variation.

Besides, considering the better arrangement of features, i.e., PC2 and PC4, it is possible to observe the performance in

computing processing. In other words, the gain in performance reaches to 30% compared to the original data.

Finally, support vector machine classifiers have been applied to classify the five different stages of growth of the Fall armyworm.

Concerning the set of SVM classifiers, the results demonstrated the efficiency and innovation of the classification method in the proposed scenario. The results also revealed that the Gaussian kernel function exhibited the best classification accuracy and precision.

The results also have shown that the developed method is capable of helping in the control of one of the main pests of maize crops, the Fall armyworm (*Spodoptera frugiperda*)

For future works, it is suggested to extend this research to an unsupervised method to reach the selection of the principal components numbers to remain with the semantic features from a recognized agricultural pest.

VI. ACKNOWLEDGMENT

This research was partially supported by the São Paulo Research Foundation (FAPESP 17/19350-2). We thank the Brazilian Corporation for Agricultural Research (Embrapa) and the Post-Graduation Program in Computer Science from the Federal University of São Carlos (UFSCar). The Authors also recognize the helpful discussions with MSc Bruno M. Moreno to finalize the manuscript.

REFERENCES

- [1] A. B. Bertolla and P. E. Cruvinel, "Dimensionality reduction for ccd sensor-based image to control fall armyworm in agriculture," in *2024 ALLSENSORS 9th International Conference on Advances in Sensors, Actuators, Metering and Sensing*. IARIA, 2024, pp. 7–12.
- [2] C. Zhang, H. Hu, D. Fang, and J. Duan, "The ccd sensor video acquisition system based on fpga&mcu," in *2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, vol. 9. IEEE, 2020, pp. 995–999.
- [3] S. Sankaran, R. Ehsani, and E. Etxeberria, "Mid-infrared spectroscopy for detection of huanglongbing (greening) in citrus leaves," *Talanta*, vol. 83, no. 2, pp. 574–581, 2010.
- [4] J. L. Miranda, B. D. Gerardo, and B. T. T. III, "Pest detection and extraction using image processing techniques," *International Journal of Computer and Communication Engineering*, vol. 3, no. 3, pp. 189–192, 2014.
- [5] C. Ji *et al.*, "Big data processing: Big challenges and opportunities," *Journal of Interconnection Networks*, vol. 13, no. 03n04, p. 1250009, 2012.
- [6] W. K. Vong, A. T. Hendrickson, D. J. Navarro, and A. Perfors, "Do additional features help or hurt category learning? the curse of dimensionality in human learners," *Cognitive Science*, vol. 43, no. 3, p. e12724, 2019.
- [7] A. L. M. Levada, "Parametric PCA for unsupervised metric learning," *Pattern Recognition Letters*, vol. 135, pp. 425–430, 2020.
- [8] F. Anowar, S. Sadaoui, and B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE)," *Computer Science Review*, vol. 40, p. 100378, 2021.
- [9] S. Ayesha, M. K. Hanif, and R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," *Information Fusion*, vol. 59, pp. 44–58, 2020.
- [10] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.
- [11] A. L. M. Levada, "PCA-KL: a parametric dimensionality reduction approach for unsupervised metric learning," *Advances in Data Analysis and Classification*, vol. 15, no. 4, pp. 829–868, 2021.
- [12] F. Pan, G. Song, X. Gan, and Q. Gu, "Consistent feature selection and its application to face recognition," *Journal of Intelligent Information Systems*, vol. 43, pp. 307–321, 2014.
- [13] S. Nanga *et al.*, "Review of dimension reduction methods," *Journal of Data Analysis and Information Processing*, vol. 9, no. 3, pp. 189–231, 2021.
- [14] Z. A. Sani, A. Shalhaf, H. Behnam, and R. Shalhaf, "Automatic computation of left ventricular volume changes over a cardiac cycle from echocardiography images by nonlinear dimensionality reduction," *Journal of Digital Imaging*, vol. 28, pp. 91–98, 2015.
- [15] R. M. Wu *et al.*, "A comparative analysis of the principal component analysis and entropy weight methods to establish the indexing measurement," *PLoS One*, vol. 17, no. 1, p. e0262261, 2022.
- [16] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Pearson Prentice Hall, 2004.
- [17] G. Saravanan, G. Yamuna, and S. Nandhini, "Real time implementation of rgb to hsv/hsi/hsl and its reverse color space models," in *2016 International Conference on Communication and Signal Processing (ICCCSP)*. IEEE, 2016, pp. 0462–0466.
- [18] W. Zhao and J. Wang, "Study of feature extraction based visual invariance and species identification of weed seeds," in *2010 Sixth International Conference on Natural Computation*, vol. 2. IEEE, 2010, pp. 631–635.
- [19] L. Zhang, F. Xiang, J. Pu, and Z. Zhang, "Application of improved hu moments in object recognition," in *2012 IEEE International Conference on Automation and Logistics*. IEEE, 2012, pp. 554–558.
- [20] K. Hongyu, V. L. M. Sandanielo, and G. J. de Oliveira Junior, "Principal analysis components: theoretical summart, application and interpretation, (análise de componentes principais: resumo teórico, aplicação e interpretação)," *E&S Engineering and Science*, vol. 5, no. 1, pp. 83–90, 2016.
- [21] B. Zhao, X. Dong, Y. Guo, X. Jia, and Y. Huang, "PCA dimensionality reduction method for image classification," *Neural Processing Letters*, pp. 1–22, 2022.
- [22] M. P. Uddin, M. A. Mamun, and M. A. Hossain, "PCA-based feature reduction for hyperspectral remote sensing image classification," *IETE Technical Review*, vol. 38, no. 4, pp. 377–396, 2021.
- [23] B. M. S. Hasan and A. M. Abdulazeez, "A review of principal component analysis algorithm for dimensionality reduction," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 20–30, 2021.
- [24] P. E. Hart, D. G. Stork, R. O. Duda *et al.*, *Pattern classification*. Wiley Hoboken, 2000.
- [25] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, 1998.
- [26] K. Faceli, A. C. Lorena, J. Gama, T. A. d. Almeida, and A. C. P. d. L. F. d. Carvalho, "Inteligência artificial: uma abordagem de aprendizado de máquina," 2021.
- [27] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, 1998.
- [28] A. C. Lorena and A. C. De Carvalho, "Uma introdução às support vector machines," *Revista de Informática Teórica e Aplicada*, vol. 14, no. 2, pp. 43–67, 2007.

Evaluating Digital Avatars - A Systematic Approach to Quantify the Uncanny Valley Effect by Using Real Life Samples

Hakan Arda

Faculty of Computer Science and Business Information
Systems
Technical University of Applied Sciences Würzburg-
Schweinfurt
Würzburg, Germany
Hakan.Arda@thws.de

Andreas Henneberger

Faculty of Computer Science and Business Information
Systems
Technical University of Applied Sciences Würzburg-
Schweinfurt
Würzburg, Germany
Andreas.Henneberger@study.thws.de

Karsten Huffstadt

Faculty of Computer Science and Business Information
Systems
Technical University of Applied Sciences Würzburg-
Schweinfurt
Würzburg, Germany
Karsten.Huffstadt@thws.de

Nicholas Müller

Faculty of Computer Science and Business Information
Systems
Technical University of Applied Sciences Würzburg-
Schweinfurt
Würzburg, Germany
Nicholas.Mueller@thws.de

Abstract—Virtual reality has seen significant advancements in recent years, particularly in rendering nearly perfect replicas of real-world objects. These improvements in environmental realism enhance user immersion. However, increasing the realism of human-like avatars seems to have the opposite effect, often leading to negative emotional reactions and breaking immersion. This is explained by the uncanny valley curve, where subtle deviations in lifelike avatars can evoke discomfort. In this study, we developed a method to evaluate human-like avatars based on the uncanny valley curve, aiming to pinpoint where this discomfort originates. A database of over 200 avatar images was created, and studies were used to identify key characteristics that make these avatars resemble humans. Additionally, we conducted a follow-up experiment with 30 participants, where real avatars were developed based on our findings and tested in live, direct conversations. This allowed us to validate our approach and offer a more precise method for future research on avatar evaluation and uncanny valley effects. Our results open up new avenues for refining avatar design, potentially mitigating negative reactions and increasing overall immersion in virtual reality experiences.

Keywords—Uncanny valley; human-like avatars; human-likeness; database; virtual reality.

I. INTRODUCTION

Based on the findings of the previous study [1], in this journal article we want to delve deeper into the Uncanny Valley and show which criteria are particularly relevant to overcome it. The Uncanny Valley effect is a psychological phenomenon that describes the unease or discomfort people experience when encountering human-like entities that are almost, but not quite, convincingly realistic. The term "Uncanny Valley" was coined by robotics professor Masahiro Mori in 1970 [2]. The concept suggests that as the appearance or behavior of humanoid entities

becomes increasingly close to human-like, there is a point at which they elicit a strong negative emotional response before eventually becoming indistinguishable from real humans.

There have already been many far-reaching attempts to investigate the effects of the human likeness of robots on people's emotional reaction [3–9]. One example of this is the work of Kim et al. [7] who used the open-source Anthropomorphic RoBOT (ABOT) database to analyze the human similarity of 251 robots. They asked a group of 150 participants to rate images of robots from the ABOT database according to their human likeness and uncanny valley factor. With the results of this survey, they have found evidence of Mori's Uncanny Valley [2]. This valley was evident in participants' perceived uncanniness of 251 robots that varied widely in terms of the range and characteristics of human likeness. They also found evidence of another, second valley of uncanniness in robots that showed a moderately weak resemblance to humans.

The developers of the ABOT [4] database took a similar approach in their study, providing a basis for research in this area. The researchers found that the human-like appearance of robots can be divided into three dimensions of human-like appearance: the robots' surface features (e.g., skin, hair, clothing), the main components of the robots' body manipulators (e.g., torso, arms, legs) and the robots' facial features (e.g., eyes, mouth, face) [7]. These results suggest that the overall perception of the physical human-likeness of robots and its relationship to emotional reactions to the robots can be explained by different constellations of the three human-like appearance dimensions. If the hypothesized uncanny valley phenomenon could be understood at the level of specific human-like appearances, this could also lead to the improvement of virtual avatars. A deeper understanding of the uncanny valley

could help developers fine-tune the design of avatars to evoke more positive emotional responses from users, enhancing user interaction and engagement. from users.

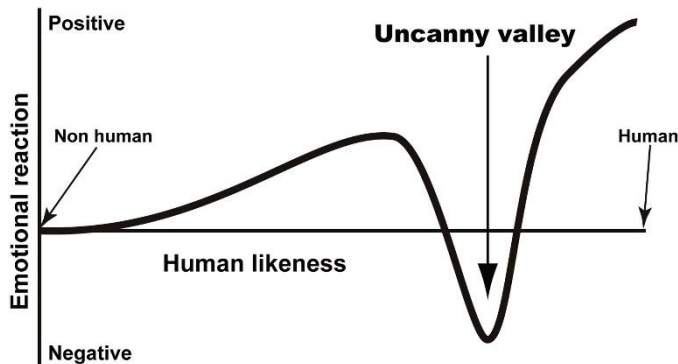


Fig. 1. Graphical representation of the Uncanny Valley. Retrieved from [10].

This approach of using a large database such as ABOT to test different human-like robots for their Uncanny Valley factors does not exist in relation to digital avatars currently. One reason for this could be the lack of such a database and the associated basis for the resulting research. Another reason is that there is currently no systematic, evidence-based approach for categorizing avatars into a continuum of perceived human-likeness. Consequently, researchers and designers are usually forced to rely on heuristics and intuitions when it comes to selecting human-like avatars for studies or developing human-like features in avatar design. This approach faces several problems. Firstly, there is currently no quantitative system to describe the degree of human likeness in different avatars, which makes it difficult to compare research results between different studies. Therefore, a precise scale is needed to compare different avatars on a common scale and allow researchers to replicate their results with their avatars.

Secondly, even when researchers manage to quantitatively assess the impression people have of an avatar's appearance, they usually treat the concept of "human likeness" as one-dimensional. However, human likeness can be expressed in different ways. For example, through gestures and facial expressions or, more generally, through the mere presence of arms and legs. Humanity has many different characteristics and therefore also different features that need to be considered.

Thirdly, the effects of the appearance of avatars on different types of avatars must be considered in the investigations. While it may be of practical advantage to limit yourself to a certain type of avatar, such as simplified human likenesses like the ones Meta uses in her Horizon. However, these restrictions can mean that certain small differences between avatars are lost. And the psychological conclusions of this work would not be applicable to a variety of other avatars.

In addition, previous studies have extensively explored methods to overcome the uncanny valley in character design. For example, the work by Schwind et al. examined how atypicality's (strong deviations from the human norm) for high levels of realism cause negative sensations in humans and

animals [11, 12]. The negative effects of atypical features, such as unnaturally large eyes or human emotions in realistic animals, are stronger for more realistic characters, than for characters with reduced realism. Consistently rendered realism can reduce the negative effects of atypicality and increase affinity, as shown by the first peak in the uncanny valley. Therefore, it is important to avoid combining realistic renderings and detailed textures of skin or eyes with non-human-like features. At high levels of realism, atypical features can cause the uncanny valley. Another possibility is to avoid so-called "dead eyes". A virtual character's eyes are crucial in determining its realism. The work of Schwind et al., which used eye tracking, found that users first fixate on the eyes before assessing other features. This is consistent with previous research showing that the realism and inconsistencies of human characters are primarily judged based on the realism of their eyes [13]. This also explains why skin makeup does not affect animacy, unlike atypical eyes or the eyes of a deceased person. The symptom of 'dead eyes' can make artificial characters feel eerie and strange. The eyes communicate intentions, behavior, and well-being, which are essential for assessing and creating affinity for the depiction.

To address these issues, and to provide a way to conduct more systematic, general, and repeatable research on virtual human-like avatars, this thesis has developed an avatar database based on the findings of Phillips et al. thus continuing the research on the uncanny valley. Furthermore, the results of this study were used to create specific categories for the creation of real avatars. In a follow-up experiment, the newly created avatars were evaluated using the same questions and criteria as in the original study, allowing for a direct comparison between the real avatars and the baseline data.

II. METHOD

A. The Avatar Database

Similar to the ABOT database by Phillips et al., the avatar database pursues three goals. Firstly, it offers an overview of the broad landscape of different human-like avatars. Secondly, the avatar database provides standardized images of human-like avatars and a growing dataset of people's perceptions of these avatars, both of which will be made public for further research in the future. Thirdly, the Avatar Database can help us better understand what makes an avatar appear human. To begin, we will discuss the development of the database. Then, two empirical studies will identify various dimensions of avatar appearance and determine which of these dimensions contribute to the perception of overall human-likeness of avatars. Subsequently, a further empirical study on the validation of the database is presented and a possibility for future research is introduced.

B. The Development of the Avatar Database

To create a comprehensive database of human-like avatar images, we searched for as many avatars as possible with the required human-like appearance characteristics. The avatars were identified from various sources, such as game characters (e.g., Fortnite, Final Fantasy, Skyrim), technology-oriented media (e.g., online magazines, newsletters, social media), companies and university websites, online communities and discussion forums, and general Google searches. To identify avatars that had not yet received significant media attention, we

also created our own avatars based on real humans and fictional characters. Between January 2024 and April 2024, we created an initial collection of over 200 images of human-like avatars, each with one or more different characteristics.

Next, we reviewed the collection of avatars and removed the ones that are already represented in the same or similar enough form. In addition, avatars that are similar to animals have also been removed because this study explicitly looks at human avatars only and uses mixed avatars later as a safety checkup. Each image also had to fulfil a certain standard to be included in the collection: No obstacles, no motion blur, no groups pictures, in color, and the entire body (With feet, hands, and head). In addition, all pictures were taken one more time in a close-up of the face. This allowed the participants to examine the entire body and then explicitly the face for the individual features. Any image that did not fulfil this standard was either not included in the collection or removed accordingly. Based on this approach we have created a collection of 200 avatars, with the corresponding source for downloading the avatars.

It was important to us that we cover as many different types as possible with the large number of avatars, such as cartoonish, stylized, realistic or minimalist. In addition, it should be possible to find an avatar that is reduced to the desired characteristics with the help of filters. This is a similar concept to searching for robots in the ABOT database. With the difference that here the avatars can be downloaded with the help of the source and used for further purposes. Some of the characters were also tested with the help of the unity engine and put to the test for suitable images. This was especially the case with avatars that we developed ourselves for this collection.

The images in the database were sorted and edited to ensure uniform recognition. This was done to ensure consistency in the images. Avatars were photographed in a frontal and neutral position with a neutral facial expression whenever possible. For avatars where this was not possible, the model was rendered again in the unity engine and photographed accordingly. Finally, the images were cropped to just the avatars with a white background using photoshop and tagged with different tags for better analysis. Here, for example, attention was paid to the recognizable gender, potential age, art style and source.

C. Measuring the Human-Likeness

In order to accurately determine the degree of humanity in avatars, the individual avatars must be evaluated according to clearly defined characteristics. Because we are dealing here with human-like avatars and not anthropomorphic robots, we are unfortunately unable to use the results of the work by Phillip et al. [4], but we can use a similar approach to determine the characteristics that we will use later. We rely on a bottom-up, feature-based approach and base our expectations on the results of the work of Phillip et al. Our goal with this approach was to define the individual features that constituted humanity in avatars and then bundle them together.

To determine the appearance characteristics of our avatars, we have developed a collection of possible characteristics that speak for humanity. We then checked all the images for the respective characteristics and deleted any that did not fit. This included features that were rare, repeated, or confusing. As a result, we defined 16 characteristics that we used for our further procedure. In addition, like Phillip et al., we decided to contribute definitions for the features. We started with relevant

definitions from the Oxford English dictionary and adapted them according to our application. For example, we were able to retain the biological functions for features such as "mouth" because our test objects are not robots but human-like avatars. This resulted in a table of features and their definitions. These definitions served as a way for our participants to focus on certain characteristics when evaluating the avatars. Since they are human-like avatars and not robots, all characteristics are always present in some form. They only differ in their design. However, there are also avatars that do not have smooth white skin and are instead green with lots of dots. Therefore, the question here is not whether the respective characteristic is present, but rather to what degree it stands out.

TABLE 1. COLLECTION OF APPEARANCE FEATURES AND ASSOCIATED DEFINITIONS

<i>Feature</i>	<i>Definition</i>
Arm	The upper limb of the human body, or the part of the upper limb between the shoulder and the wrist.
Eye	The organ of sight. Either of the paired globular organs of sight in the head of humans.
Eyebrow	The (usually arched) line of short fine hair along the upper edge of each of a person's eye sockets.
Eyelashes	The line of hairs fringing each edge of an eyelid, serving to help keep the eye free of dust or other extraneous matter.
Face	The front part of the head, from the forehead to the chin, and containing the eyes, nose, and mouth.
Finger	Each of the five slender jointed parts attached to either hand.
Genderedness	Features of appearance that can indicate biological sex, or the social categories of being male or female.
Hand	The terminal part of an arm, typically connected to the arm by a wrist.
Head	The uppermost part of a body, typically connected to the torso by a neck. The head may contain facial features such as the mouth, eyes, or nose.
Head hair	A collection of threadlike filaments on the head.
Leg	The lower limb of the human body, or the part of the lower limb between the hip and the ankle.
Mouth	The orifice in the head of a human or other vertebrate through which food is ingested and vocal sounds emitted.
Nose	The part of the head or face in humans which lies above the mouth and contains the nostrils.
Skin	The layer of tissue forming the external covering of the body.



Fig. 2. All 200 human-like vr avatars in the database.

D. Measuring the Uncanniness

To address the question of how uncanny the avatars appeared, we followed the methodology outlined in the study by Kim et al. [7]. Their research involved a large-scale investigation utilizing the ABOT database, building on the foundational findings of Phillip et al. [4]. Kim et al. conducted a follow-up study that specifically compared robots based on their levels of human-likeness and perceived uncanniness. Drawing inspiration from their approach, we adopted the same definition of uncanniness to ensure consistency and standardization in our evaluation process. To help participants understand and apply a uniform criterion during the evaluation, we provided them with a clear definition of uncanniness. Specifically, they were informed that uncanniness refers to "the characteristic of seeming mysterious, weird, uncomfortably strange, or unfamiliar." This definition was derived directly from the Oxford English Dictionary's explanation of the term, ensuring its alignment with established and conceptual standards. By using this definition, we aimed to create a common framework to assess the avatars' uncanniness systematically.

E. Real Life Avatar Samples

For the creation, certain criteria were considered that were based on the results of the first study. These categories are mentioned in the results section.

There are numerous methods for the creation of a lifelike or stylized digital persona. One method is to scan the user in their entirety in order to create a lifelike avatar (Fig. 3, far right). The process of creating virtual avatars through scanning involves capturing the physical features of real individuals and translating them into digital representations. This process is primarily

reliant on the utilization of cameras and bespoke software in order to achieve accurate and realistic virtual avatars. In the initial stage of the process, an iPhone 13 Pro camera and the Polycam application were employed. The camera of the iPhone 13 Pro is equipped with the requisite technical capabilities for the capture of 3D scans. In the data capture phase, this camera is employed to record information from the subject. The cameras of the iPhone 13 Pro range from conventional RGB (red, green, blue) cameras to advanced depth-sensing cameras such as LiDAR (light detection and ranging) or structured light cameras. Their function is to capture a range of characteristics of the individual, including their appearance, shape, and texture.

Depth-sensing cameras are of particular significance, as they provide data regarding the spatial distance between the camera and various points on the individual's body. This depth data serves as the basis for the construction of a three-dimensional representation of the individual. Concurrently, RGB cameras capture color data, which is indispensable for texturing the 3D model and endowing it with visually realistic details. To ensure accuracy, the cameras are calibrated and aligned. Calibration ensures that the cameras are correctly configured and calibrated, whereas alignment determines the precise positions and orientations of the cameras in relation to one another. Such meticulous calibration is of paramount importance for ensuring the seamless integration of data from disparate camera perspectives. The subsequent phase is to utilize the data obtained from the camera in the Polycam application. The application utilizes the data to generate a "point cloud," which is a set of 3D points in space that collectively delineate the surface of the person.



Fig. 3. Real life avatar samples. From left to right; the generic, - stylized, - mixed, - and lifelike avatar.

This point cloud represents the initial representation of the person's physical form. Subsequently, the point cloud is transformed into a "mesh", which is a network of interconnected triangles that collectively create a coherent three-dimensional surface. The resulting mesh provides a more detailed and accurate representation of the person's body. In order to imbue the virtual avatar with realistic visual details, texture information derived from the RGB images is projected onto the mesh, thereby imparting the avatar with characteristics such as skin color, clothing patterns, and other visual attributes. In order to optimize the final result, the mesh was subjected to a cleaning and refinement process. The objective of these steps is to eliminate imperfections, smooth surfaces, and eliminate any undesired artefacts that may have emerged during the earlier stages.

In order to facilitate dynamic poses and movements, it is necessary to provide the virtual avatar with a skeletal structure. The process of rigging involves the creation of a digital skeletal structure with joints and bones, which emulates the human anatomy. This stage was completed using the Mixamo platform. This software is a free product from Adobe and offers a plethora of animations, avatars, and possibilities for rigging. In rigging, the software is employed to construct a skeleton or rig of bones or joints that determines the manner in which the constituent parts of a mesh can be moved. The resulting skeleton is then connected to the avatar, thus enabling the latter to perform animations and poses that mirror real-world actions. Once the avatar has been rigged, it becomes possible to animate it through various methods, including the utilization of motion capture data or virtual reality headsets. This animation process imbues the avatar with a lifelike quality, enabling it to replicate the movements and expressions of the real person on whom it

was based. The subsequent stage is to incorporate the avatar into the Unity environment.

In this instance, the avatar is furnished with an animation controller. The controller oversees the transitions between disparate animations, thereby enabling the user to regulate the animations in accordance with parameters such as movement, gestures, and interactions. In order for the headset to transmit these parameters, the head mounted display (HMD) must be integrated into Unity with the assistance of a VR software development kit (SDK). A variety of HMD manufacturers, including Meta, HTC Vive, and Steam, provide their own SDKs. However, there is also an open-source variant, OpenXR, which promises compatibility across all headsets. Integration of OpenXR into existing projects is a straightforward process, requiring only configuration for the specific headset in question. Accordingly, OpenXR was employed in this study. The SDK serves as an interface between the hardware and the software, routing all parameters to the requisite scripts of the animation controller and the HMD. Ultimately, the individual body parts of the avatar must be connected to the corresponding hardware controllers. With slight adjustments, the digital persona can be readily exported and imported between various scenes in Unity.

F. Stylized Avatar

An additional method for creating a more stylized avatar (Fig. 3, second from left) is through the utilization of the tool Ready Player Me. The process of creating an avatar using the Ready Player Me tool is relatively straightforward [14]. To commence, open your web browser and navigate to the relevant website. On the website, a button or link will be provided, which will prompt the user to proceed with the creation of their avatar. Once the process has been started, the initial step is to

specify the gender of the avatar. A series of options will then be presented, including the choices of male, female, and other gender options. After the gender has been selected, the next step is to proceed with the customization of the avatar. This entails making adjustments to a number of features in order to ensure that the avatar's appearance aligns with your personal preferences. The available customization options may include the selection of a hairstyle, hair color, modification of facial features, choice of clothing items, and potentially other features. By selecting the requisite customization option, the desired adjustments can be made to create an avatar that reflects the user's envisioned appearance. Once the avatar has been meticulously customized, the finalization stage will be reached. At this point, the user will encounter an option labelled "Finish" or "Export." This is the phase during which the system generates a distinctive avatar based on the user's selections. Upon completion of the aforementioned steps, the avatar will be available for download. The format in which the avatar is provided will depend on the intended use.

One advantage of Ready Player Me is that no additional hardware is necessary, with the exception of a device with web browsing capabilities. The avatar is equipped with all the requisite elements, including mesh, textures, and a skeleton, immediately following the download. It can be operated directly with the HMD. However, the lack of photographic input limits the degree of realism achievable with the avatar. In contrast, the process here is based on the user's perception of themselves. It is possible to create an avatar based on a photograph. In order to utilize this functionality, it is necessary to capture an image of the user's face. Subsequently, the program identifies the most salient features of the face and searches the database of potential "bricks" to determine the optimal match. Such characteristics include hair color and style, eye color, eyebrows, nose, mouth and general facial shape. Subsequently, the user is afforded the opportunity to personalize the avatar they have created, selecting from a range of physiques, clothing and accessories.

G. Mixed Avatar

To offset this issue, a third method for creating avatars has been developed, which combines the benefits of lifelike and stylized avatars to create a hybrid (Fig. 3, third from left) that encompasses both approaches. In the case of avatars intended to closely resemble a real person, the face is likely to be of paramount importance. The face is a reflection of the individual's behavior, empathy, and perception [15, 16]. The website Avaturn employs this characteristic, offering avatars that are created expeditiously yet remain highly realistic [14]. To achieve this, the software employs the same methodology as that used by Polycam for scanning individuals, but restricts its application to the user's face. This necessitates the capture of three biometrically-verified images of the user's face. Three photographs are required, taken from the front, the left, and the right side, respectively. The photographs serve as the basis for the creation of a digital mesh and texture for the face. Subsequently, the remaining body parts, including hair, the torso, the arms, and the legs, are created using a character creator tool that is similar to the Ready Player Me tool. At the

outset, the user is prompted to select their gender. Subsequently, the user is afforded the opportunity to modify the height and stature with the assistance of sliders. Afterwards, the user may select from a range of hairstyles, clothing, and accessories. Additionally, the software provides the option to incorporate animations and other design elements, thus allowing users to customize their avatar to a considerable extent. Furthermore, as the body is constructed from a predefined model, the avatar can be imported directly into Unity and linked to the animation controller, in a manner analogous to that observed in the Ready Player Me avatar. The result is an avatar that, upon initial observation, appears to be a near-identical replication of the original. However, upon closer examination, notable discrepancies emerge. This is not, however, an issue in itself, as an avatar that is too similar to humans can be subject to the uncanny valley effect, which can result in a reduction in immersion [17].

Furthermore, an avatar devoid of any connection to the human subject was included as a control group for the tests. The character model in question is that of Space Robot Kyle (Fig. 3, far left). This robot, which is devoid of any distinctive characteristics, provides all the necessary features to create a VR avatar, without specifying a gender or body shape. As the robot does not aspire to emulate the human form, the potential for evoking the Uncanny Valley effect is minimal [2]. Consequently, it is feasible to ascertain whether there are discernible discrepancies between realistic and generic avatars. The four types of avatars, namely lifelike, stylized, mixed and generic, constitute the fundamental variable upon which the experimental design of the second case study is based.

H. Participants Study 1

For the first study, we recruited a total of 160 participants via Prolific crowdsourcing website. Data collected via crowdsourcing websites such as Prolific is currently very much in vogue. This is mainly due to the fact that the data can be collected very easily and quickly and there are already studies showing that the data collected here can keep up with traditional methods in terms of quality [18]. Nevertheless, the data should also be checked for quality [19]. We have therefore decided to incorporate various quality checks into the data collection process. Firstly, all data sets with incorrect answers to six or more "catch trials" are removed. Secondly, we considered a lack of variation in ratings between participants as an indicator of inattention. Therefore, we removed data from participants whose ratings had a standard deviation of less than 10 ($SD < 10$) on a scale of 0 - 100. Finally, we compared each participant's ratings with the average of the remaining judgements in their group (between participants) by calculating the correlation between the individual judgements and the remaining judgements in their group. If this correlation between the individual participant's ratings and the group mean was less than 30, the participant's data were discarded as these individuals may have been performing a different assessment task to the group. After this quality check, the total number of participants, from around the world, was 143 (M Age = 20, SD Age = 10, 104 Male, 41 Female, 2 No Responses). This means that each avatar had a rating of 15 - 20 participants.

I. Design and Procedure Study 1

The 200 avatars were divided into four groups of 50 avatars each. Each group was also provided with 10 catch trials. This meant that each group had 60 avatars, which were rated by 20 participants. Because two different questions were asked in this study, one asking, "how human-like is the avatar?" and the other "how uncanny is the avatar?", we asked each group twice. This gave us a total of 4 groups of 60 avatars per question.

The participant begins the survey with a brief introduction to the topic and a short briefing on how to complete the survey. This was followed by an example task on how the participant should rate the avatars. The same example avatar was used for each block. The participant sees two pictures of an avatar on their screen. On the left the entire body and on the right the profile picture with the face in focus. Below the pictures is the definition of the respective question. For the question about human likeness, the participant sees the various characteristics that make an avatar like humans and a slider from 0 to 100. Above the slider is the question "How similar to humans do you think this avatar is?" (0 - not at all like humans and 100 - very similar to humans). We used a similar method for the question of how uncanny you think the avatars are. You can see the same pictures and again a slider from 0 to 100 but this time with the definition about uncanniness and the question "How uncanny do you think this avatar is?" (0 - not uncanny at all and 100 - very uncanny). The participants are randomized into one of the respective groups and are only allowed to answer one question type. This is to prevent the questions from influencing each other. The catch trials are pictures of real people or objects that have also been randomized into the respective groups.

After half of the questions, the participants were given a 10-second break during which their attention was drawn to the definition and characteristics again. After judging all the images, participants were asked to complete a demographic questionnaire in which they were asked to indicate their age, gender, native language, level of education, previous knowledge of robotics and experience with virtual avatars. The entire study took approximately 5 minutes to complete, and participants received \$1 as compensation for their participation.

J. Participants Study 2

A total of 30 participants took part in the second study. The experimental group is comprised of 29 male and one female participant. The participants ranged in age from 16 to 50, with the majority falling between the ages of 20 and 28. The probands themselves rated their experience with VR/AR on a scale of 1 (no experience) to 5 (high experience). The mean score was 2.20. The average test time was 57 minutes and 10 seconds. This figure includes the time taken to conduct the interviews and complete the survey. The majority of participants had a background in education. The study group comprised 15 bachelor's and master's students, 4 computer scientists, and 6 individuals from other non-informatics professions. It is reasonable to assume that a computer scientist or prospective computer scientist in college will have some experience with VR or AR. However, it is not possible to derive any concrete findings from this, given that each course of study has a different focus and the topic of VR and AR was not included in the subjects' basic studies. Accordingly, the

experience values assigned to the probands represent an approximate estimation of their own interest and experience. In contrast to the initial study, the participants were not sourced through crowdfunding platforms such as Prolific. Instead, they were recruited through the university, in person, through advertising, or through personal connections. Except for a few small items such as sweets or pens, there was no monetary compensation for the subjects of the second study.

K. Design and Procedure Study 2

The second study consists of three steps. First, the participants are informed about the upcoming events. It is explained that this is a role play in which the respondent takes the role of a politician and the interviewer takes the role of various journalists from different newspapers. The two parties then meet in a generic virtual space provided by the university. The respondent has a generic avatar with no particular resemblance. His counterpart is one of four avatars: lifelike, stylized, mixed, or generic.

The second step is the beginning of the discussion. The journalist asks if the politician knows the last generation. The last generation is a group of climate activists who have gained attention for their direct-action protests aimed at raising awareness about climate change. Depending on the answer, either a short explanation based on an example of the last generation or a direct question about the politician's opinion on the topic follows. Based on the politician's opinion, an argument for or against the last generation follows. The politician can respond to this argument. The journalist then proposes a possible solution to the problem and discusses whether it is within the politician's possibilities. This discussion continues until a compromise is reached or the attempt to find one is declared a failure by both parties. After the discussion, there is a short break during which the interviewee can answer the first two questions regarding the uncanniness and the human likenesses. The same questions and definitions were used here as in the first study. The interview takes place four times in total. Each time, the avatar's appearance and newspaper will change. After each round, the participant has time to answer the questions. After the fourth round, the discussion ends and the final phase begins.

In the last phase, the respondent is informed about the topic and asked to fill in the voluntary feedback questions about the avatars. After that, there is a farewell and the test is successfully completed.

III. RESULTS

The findings of the initial and subsequent studies are presented individually in the following sections, as the latter was conducted post-hoc.

A. Results Study 1

For the data analysis, all the results of the individual surveys were added together and an average for human likeness and uncanniness was calculated for each avatar. We then inserted these results into Microsoft Excel to generate various graphs. Looking at the first graph (Fig. 3 a), no uncanny valley can be recognized. Instead, there is a linear gradient between the two

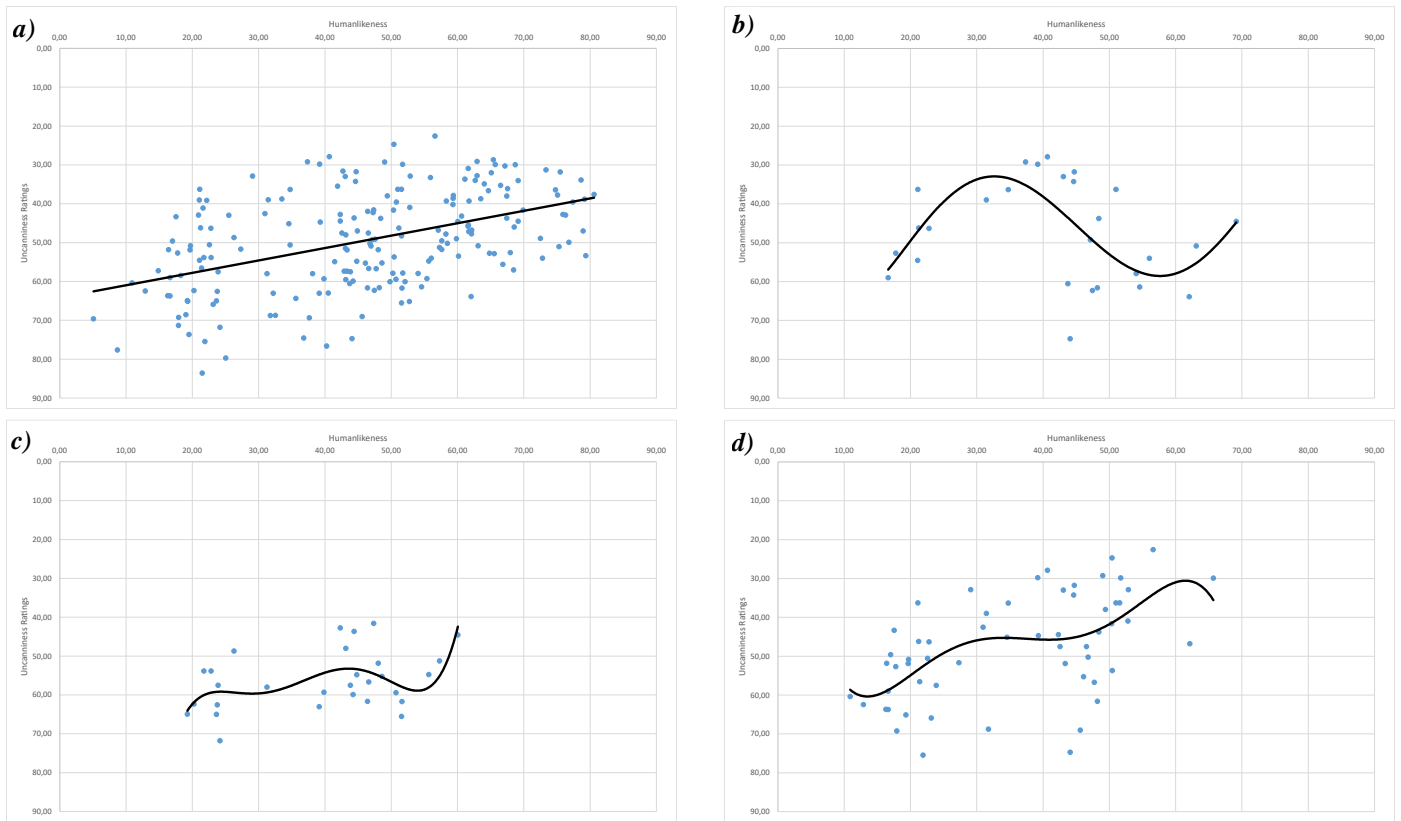


Fig. 4. Four scatterplots a) Total scope of all avatars b) Only avatars representing children c) Avatars caricaturing a real-life person d) Avatars representing a cartoonized style. This Y axis has been inverted to represent the uncanny valley.

factors uncanniness and human likeness, with uncanniness decreasing as human likeness increases. However, if one does not look at the entire amount of data and instead only at certain categories, such as only avatars that represent children, an uncanny valley is clearly recognizable. Just as in Mori's uncanny valley hypothesis [2], a large valley can be recognized between the moderately realistic and realistic avatars. When looking at other avatar categories, a slight uncanny valley can also be recognized. The other avatar categories, such as avatars that are based on a real person and represent them as a caricature, also have a slightly uncanny valley. The same applies to avatars that are not based on a real person but are depicted as a cartoon. based on a real person and represent them as a caricature, also have a slightly uncanny valley. To further investigate these results, we performed a polynomial mixed fit for the three different categories of. We determined the different coefficients of determination = r^2 for different polynomial mixed effects 3rd, 4th, and 5th models. In addition, based on the results of Kim et. al. [7] we also assumed that if there are one or more valleys here, then these are recognized in the 4th or 5th polynomial model.

B. Results Study 2

As with the initial study, the findings of the individual surveys were aggregated and imported into Microsoft Excel. Subsequently, based on the preceding graphs, structural graphics were generated in this manner as well.

The new graph (Fig. 5) illustrates the ratings of the four new avatars (generic, stylized, mixed, and life-like) in comparison to the other avatars based on their similarity to humans and

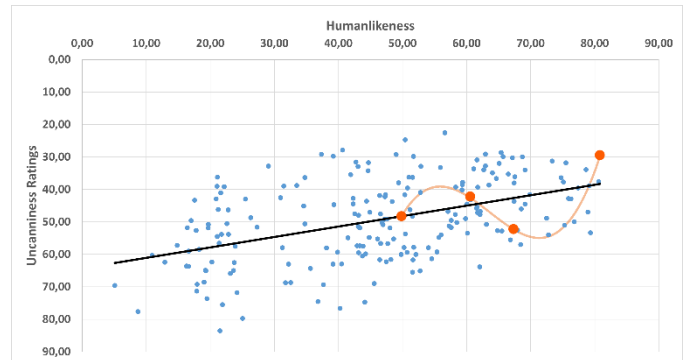


Fig. 5. Scatterplot of all avatars including new generic, stylized, mixed, and life-like avatars in order from left to right.

perceived uncanniness. For enhanced readability, the four new avatars were marked in red and 200% larger than the other points on the graph. The generic avatar (Fig. 5, far left) scored 49.87 points in human likeness and 48.20 in uncanniness. This places the avatar in a relatively intermediate position in comparison to other avatars. The stylized avatar (Fig. 5, second from left) is more human-like than the generic avatar, with a higher score of 60.47 on the human likeness and a slightly lower score of 42.50 on the uncanniness scale. When the data set is considered as a whole, the curve demonstrates that the degree of uncanniness decreases as the degree of human likeness increases. This finding is comparable to the results observed in Fig. 4 a). The aforementioned increase is, however, interrupted by the mixed

avatar (Fig. 5, third from left), which, despite its greater similarity to humans with 67.33 points, was rated as more uncanny with 52.30 points. Furthermore, the mixed avatar exhibits a significantly lower mean than the overall distribution of all avatar in the same human likeness category. The life-like avatar (Fig. 5, far right) is deemed the most human-like, with an average score of 80.81 on the human likeness scale and 29.81 on the uncanniness scale. This makes the avatar the most human-like and least uncanny of the new avatars, and one of the best performing overall. As a result, an uncanny valley curve emerges between the new avatars, although this is of limited significance due to the small sample size of four avatars. Accordingly, the r^2 values were not included in the analysis. The following chapters will present a discussion of the results, together with a comparison to the findings of the initial study.

IV. DISCUSSION

Using a database of 200 different VR avatars, we were able to find evidence for the uncanny valley phenomenon of Mori et al. [2]. Contrary to expectations, however, this was not the case for the entire sample, but only when we looked more closely at different sub-categories of avatars. This means that when many different avatar categories, such as different age classes or different styles, are analyzed together, the individual graphs overlap and thus close the uncanny valley. The valley can only be created if the data is sorted precisely.

Particularly noticeable here were avatars representing children. We found that when trying to make this type of avatar more realistic, some avatars were perceived much more negatively than avatars that did not try. Avatars that received a lower human-likeness score of under 40 out of 100 points for uncanniness were significantly better than those with a higher humanlike score and over 60 out of 100 points for uncanniness. This phenomenon cannot be replicated in the other age groups. We assume this is due to the proportions of the avatars. Because the unrealistic avatars in particular have a significantly larger head than the realistic avatars, which have normal proportions here. We were able to make this observation with the caricatures of real personalities such as former presidents of the USA. An uncanny valley can also be recognized here and, similar to the avatars representing children, these are mainly avatars with unusual proportions. This could be due to the fact that an attempt was made here to depict real people and by increasing the similarity the avatars fall into uncanny again.

By focusing solely on the uncanny valley phenomenon in each curve, it is now possible to hypothesize certain characteristics that the avatars possess in common with one another. For example, all the avatars represented on the uncanny valley curve exhibit the "dead eye" syndrome [13]. As the face and eyes are the most prominent features, it is unsurprising that these avatars are perceived as more uncanny. It is likely that this is also the reason why the mixed avatar performed significantly worse than the other avatars (Fig. 5). This is because, when the avatar's eyes are generated, the actual textures of the eyes are not taken from the photograph; instead, only an approximation is generated by the computer, which results in the eyes appearing lifeless (Fig. 3, third from left). Furthermore, a significant number of the avatars identified as uncanny exhibit disproportionate limb sizes. In most cases, the head is the area of the avatar that is perceived as disproportionate. However, the

shoulders and hands are also often observed to be out of alignment with the rest of the body. This may explain why the children and caricature avatars from Fig. 4. b) and c) have been classified in such a diverse manner. It is challenging to find the appropriate proportions for avatars representing children without appearing unnatural or unsettling. To "cute up" a human usually means to adjust the proportions so that the head is larger than the rest of the body. Similarly, the avatars caricature famous personalities. Here, too, larger heads were used to make the person behind the personality appear cute or amusing. Another feature is the skin of the avatars. Particularly striking were the avatars that, although they were humanoid in stature, had little or no skin similar to that of humans.

To illustrate, the avatar with the highest degree of uncanniness among the simple avatars was the "Mannequin" avatar (Fig. 6). Furthermore, this avatar exhibits the highest degree of uncanniness when evaluated across the entire database.



Fig. 6. The image of the mannequin, which has the highest score for uncanniness.

Furthermore, the aspect of poor skin quality also serves to differentiate the avatars, which exhibited a similarly elevated level of uncanniness. Mori had previously referenced this category in his uncanny valley curve, situating these types of avatars at the lowest point on the curve [2]. The aspects that have been identified as the most significant contributors to the perception of uncanniness can be distilled into three primary criteria: eyes, body proportions, and skin. By incorporating these three criteria into the design of avatars, it is possible to create representations of individuals that are not perceived as uncanny, as shown by the stylized (Fig. 3, second from left) or lifelike avatar (Fig. 3, far right) in the second study (Fig. 5).

V. CONCLUSION

With the drastic development of virtual reality and the constantly growing environment and possibilities it offers us, human-like avatars are also becoming an important topic that will affect us in the coming years. Even now, avatars from different areas are being rated according to their appearance and the term uncanny valley is being used more and more frequently. Based on the results of this study, we were able to find out that the uncanny valley is not an all-encompassing

phenomenon in relation to VR avatars. Instead, the uncanny valley can only be found when taking a closer look at the individual subcategories of avatars. For example, if you take all the avatars in this database, there is an increasingly linear development between human-likeness and uncanniness, with the uncanny factor decreasing as human-likeness increases. However, if you look at certain subcategories, you can see a valley. This observation can also be observed in other categories, which leads us to assume that an overlap between the individual categories means that the uncanny valley is closed and thus balanced out by different avatars.

A more detailed examination of the uncanny valley revealed three primary factors: eyes, body proportions, and skin. These factors, among others, exert a significant influence on the perception of uncanniness. These factors were used to create actual example avatars, thereby enabling the generation of realistic representations of individuals who exhibited minimal or no uncanniness. Consequently, it can be reasonably inferred that with the assistance of enhanced scanning options and adherence to specific criteria, it is feasible to create avatars that accurately reflect the human form. In order to confirm this assumption, further and possibly even larger-scale studies than this one are needed. And by continuing to develop this database, we want to make this possible for everyone.

ACKNOWLEDGMENT

This work was supported by the Würzburg-Schweinfurt University of Applied Sciences. We would like to acknowledge the work of Kim et al. and the developers of the ABOT database. Their work can be found in [4] and [7]. Furthermore, we would like to thank everyone who participated in the study, especially those who participated in the second study, which took a relatively long time to conduct.

REFERENCES

- [1] Arda H, Henneberger A. Evaluating Digital Avatars in VR - A Systematic Approach to Quantify the Uncanny Valley Effect. In: Stigberg S, Karlson J, Müller NH, editors. ACHI 2024: The Seventeenth International Conference on Advances in Computer-Human Interactions : May 26-30, 2024, Barcelona, Spain. Wilmington, DE: IARIA; 2024. p. 1–6.
- [2] Mori M, MacDorman K, Kageki N. The Uncanny Valley [From the Field]. *IEEE Robot. Automat. Mag.* 2012;19:98–100. doi:10.1109/MRA.2012.2192811.
- [3] Verplank B, Sutcliffe A, Mackay W, Amowitz J, Gaver W, editors. Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques. New York, USA; 2002.
- [4] Phillips E, Zhao X, Ullman D, Malle BF. What is Human-like? In: Kanda T, Šabanović S, Hoffman G, Tapus A, editors. HRI '18: ACM/IEEE International Conference on Human-Robot Interaction; 05 03 2018 08 03 2018; Chicago IL USA. New York, NY, USA; 2018. p. 105–113. doi:10.1145/3171221.3171268.
- [5] Mathur MB, Reichling DB. Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition.* 2016;146:22–32. doi:10.1016/j.cognition.2015.09.008.
- [6] MacDorman KF, Ishiguro H. The uncanny advantage of using androids in cognitive and social science research. 2006;7:297–337. doi:10.1075/is.7.3.03mac.
- [7] Kim B, Bruce M, Brown L, Visser E de, Phillips E. A Comprehensive Approach to Validating the Uncanny Valley using the Anthropomorphic RoBOT (ABOT) Database. In: 2020 Systems and Information Engineering Design Symposium (SIEDS); 24.04.2020 - 24.04.2020; Charlottesville, VA, USA: IEEE; 2020. p. 1–6. doi:10.1109/SIEDS49339.2020.9106675.
- [8] DiSalvo CF, Gemperle F, Forlizzi J, Kiesler S. All robots are not created equal. In: Verplank B, Sutcliffe A, Mackay W, Amowitz J, Gaver W, editors. DIS02: Designing Interactive Systems 2002; 25 06 2002 28 06 2002; London England. New York, NY, USA: ACM; 2002. p. 321–326. doi:10.1145/778712.778756.
- [9] Bartneck C, Kanda T, Ishiguro H, Hagita N. Is The Uncanny Valley An Uncanny Cliff? In: RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication; 26.08.2007 - 29.08.2007; South Korea: IEEE; 2007. p. 368–373. doi:10.1109/ROMAN.2007.4415111.
- [10] Sasaki K, Ihaya K, Yamada Y. Avoidance of Novelty Contributes to the Uncanny Valley. *Front Psychol.* 2017;8:1792. doi:10.3389/fpsyg.2017.01792.
- [11] Chattopadhyay D, MacDorman KF. Familiar faces rendered strange: Why inconsistent realism drives characters into the uncanny valley. *J Vis.* 2016;16:7. doi:10.1167/16.11.7.
- [12] Schwind V, Wolf K, Henze N. Avoiding the uncanny valley in virtual character design. *Interactions.* 2018;25:45–9. doi:10.1145/3236673.
- [13] Moore RK. A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. *Sci Rep.* 2012;2:864. doi:10.1038/srep00864.
- [14] Ready Player Me. Integrate an avatar creator into your game in days - Ready Player Me. 13.09.2024. <https://readyplayer.me/de>. Accessed 13 Sep 2024.
- [15] Forsythe SM. Effect of Applicant's Clothing on Interviewer's Decision to Hire. *J Applied Social Psychol.* 1990;20:1579–95. doi:10.1111/j.1559-1816.1990.tb01494.x.
- [16] Zebrowitz LA, Montepare JM. Social Psychological Face Perception: Why Appearance Matters. *Soc Personal Psychol Compass.* 2008;2:1497. doi:10.1111/j.1751-9004.2008.00109.x.
- [17] Ding LI, Moon H-S. Uncanny Valley Effect in the Animation Character Design - focusing on Avoiding or Utilizing the Uncanny Valley Effect. *Cartoon and Animation Studies.* 2016;43:321–42. doi:10.7230/KOSCAS.2016.43.321.
- [18] Mortensen K, Hughes TL. Comparing Amazon's Mechanical Turk Platform to Conventional Data Collection Methods in the Health and Medical Research Literature. *J Gen Intern Med.* 2018;33:533–8. doi:10.1007/s11606-017-4246-0.
- [19] Ahler DJ, Roush CE, Sood G. The micro-task market for lemons: data quality on Amazon's Mechanical Turk. *PSRM.* 2021:1–20. doi:10.1017/psrm.2021.57.

AI-based Automated Production of Learning Content – A Means to Bridging the Digital Divide in Workplace Learning?

Katharina Frosch

Brandenburg University of Applied Sciences
Brandenburg an der Havel, Germany
email: frosch@th-brandenburg.de

Friederike Lindauer

Brandenburg University of Applied Sciences
Brandenburg an der Havel, Germany
email: lindauer@th-brandenburg.de

Abstract—This article examines the digital divide in workplace learning, highlighting disparities in the distribution and adoption of advanced learning technologies across workplace types. Through a rapid literature review, the study finds a concentration of advanced learning technologies use in the education, health, and medical sectors, with limited use in other sectors, particularly non-technical and smaller organizations. The findings underscore the inequitable access to technology-enabled learning opportunities, highlighting a significant gap between professional and non-technical sectors. The article proposes the use of AI-driven content creation as a strategic approach to democratize workplace learning and reduce the resource barriers associated with implementing advanced learning technologies. This research establishes a foundation for understanding and addressing the digital learning divide, and suggests future directions for more equitable technology integration in workplace learning.

Keywords—*technology-enhanced learning; digital divide; workplace learning; generative AI; automated content generation*

I. INTRODUCTION

Measuring and combatting the digital divide in workplace learning is of crucial importance [1]. If we fail to address this issue, the gap between those who have access to advanced learning technologies (ALT) and those who do not will widen, leading to significant disparities in skills and opportunities. This threat results from the fact that a large share of lifelong learning occurs during and alongside work and often has a rather informal character [2], [3]. In this context, much is foreseen from ALT. ALT are characterized by careful instructional design, a high degree of interactivity, and a holistic approach to the assessment of learning outcomes [4]. Some examples of ALT are adaptive learning systems, mobile micro-learning, augmented or virtual reality applications, and even digitally supported types of collaborative ("social") learning.

When designed well, these technologies can make self-regulated learning-on-the-go at the workplace easier, allowing individuals to take control of their learning and regulate it according to their needs [4]-[6]. However, if the digital divide in workplace learning is allowed to persist or even increase, those without access to these technologies will be at a severe disadvantage. They may struggle to keep up with the rapid pace of technological advancements and evolving job requirements, ultimately hindering their professional growth and career prospects. Therefore, it is essential to ensure

equitable access to ALT to boost lifelong learning for all employees, thereby fostering a more inclusive and skilled workforce.

Against this background, the first objective of our research is to gain insights into the relative distribution of opportunities to benefit from ALT for workplace reskilling and upskilling, i.e., the digital divide in workplace learning. In the second step of our research, we examine the underlying causes of the digital divide in workplace learning. Our analysis suggests that while barriers to technology adoption may play a role, they are not the sole factor contributing to the digital divide in workplace learning. In contrast, it can be observed that the substantial financial investment required to develop effective digital learning content for the workplace may represent a significant obstacle for organizations seeking to provide digital learning opportunities to their employees. It is therefore reasonable to consider whether generative AI could provide a solution to this problem, potentially reducing the time required to produce digital learning content.

In conclusion, we argue that a deeper understanding of the digital divide in workplace learning, and how AI-based content creation could mitigate it, could be an important step towards more equitable access to ALT, facilitating personal and professional growth, employability, and thus the advancement of social justice and inclusion. The following section outlines the structure of the paper. Section II presents a rapid literature review on the digital divide in workplace learning. Section III evaluates the extent to which the digital divide may be attributed to barriers related to the production of learning content at the workplace, and how AI could facilitate the generation of learning content at the workplace, thus narrowing the digital divide in workplace learning. Section IV concludes and provides some directions for further research.

II. THE DIGITAL DIVIDE IN WORKPLACE LEARNING

A. State-of-the-Art

In the past, inequalities in access and use of Information Technology (IT) have been discussed against the backdrop of the concept of the "digital divide", i.e., "digital inequalities between individuals, households, businesses or geographic areas" that arise from disparities in physical access to IT infrastructures, digital competency of users but also in unequal capabilities, engagement, and use outcomes [8]. So far, the digital divide has been, for example, discussed at the individual (i.e., age, income, educational level, digital

competencies, language barriers) level and the regional level (country, remote areas vs. rural areas) [9]. During the COVID-19 pandemic, we have experienced firsthand that the digital divide can severely limit access to education for those who are digitally left behind [9]-[12], leading to reduced education equity [14]. To our knowledge however, there is no systematic analysis yet that sheds light on the digital divide in *workplace* learning, i.e., processes related to learning and training activities at various levels of an organization, thus *at work* [14][15].

For this paper, and drawing on the general definition of the digital divide provided by [8], we define the digital divide in workplace learning by the variations in the utilization and adoption of adult learning practices across different types of workplaces. More concretely, we hypothesize that whether one works in a small or a large company, whether one works in the public or the private sector, and what job field (e.g., blue vs. white collar) one is working in severely affects one's opportunities for technology-enhanced learning. From a workplace ethics and sustainable development perspective, access to opportunities for re- and upskilling From the perspective of workplace ethics and sustainable development, access to lifelong learning opportunities should not depend on job characteristics, but should be inclusive and equitable, as required by the United Nations Sustainable Development Goals [17]. Furthermore, barriers in the access to ALT at the workplace create disparities for individual workers and puts the up- and reskilling of our workforce at risk, which is urgently needed for future employability.

Earlier studies show that the use of ALT is heavily skewed towards the educational sector [17][18], as well as towards academic professions, in particular health and medical care (ibid.) and information technology [19][20]. To give an example, in the review study by Granić [18], about 80 percent of the studies covered came from the educational field. Similarly, in the review by Yu et al. on information technology in workplace learning [21], 18 out of the 60 studies analyzed were from the medical field. There is also some evidence that ALT is less used in public services (3 out of 60 studies in the review of Yu et al. [21]) than in business enterprises [20][21] – 3 as compared to 34 in the review by Yu et al. [21] – and that smaller and medium-sized enterprises lag behind in the adoption of ALT [23].

However, even if the studies mentioned above provide informative starting points, we argue that a reliable and more granular picture of the digital divide in workplace learning is missing: Most studies rely primarily on evidence predating 2020, before the digitization boost caused by the COVID-19 pandemic. Therefore, they can be considered somewhat outdated. Two of the three studies covering very recent evidence do not [19] or not fully [23] qualify as *systematic* reviews. Recent systematic reviews cover rather specific topics such as instructional planning in e-learning [24] or the effect of technology-enhanced learning and training on organizational-level learning outcomes [20], or they focus on specific occupations and sectors, in particular those such health professionals [25] or teachers [26] where the use of ALT is frequent. The most recent systematic review by Yu et al. [21] found that only 19 out of the 60 studies analyzed (ibid,

p. 4912) focused on individual employee learning processes within enterprises. The remaining studies investigate the interplay of meta-constructs, such as technology acceptance of ALT in general or satisfaction with online forms of learning at the workplace rather than focusing on individual-level workplace learning processes. However, the review does not provide a detailed analysis of institutional characteristics or delve deeply into ALT. The current literature highlights how little we know about the varied utilization of ALT across industries, occupations, and diverse institutional settings (e.g., large vs. small, public vs. private).

To address the described gap in the literature, we propose an alternative approach to analyzing the literature on technology-enhanced workplace learning. We advocate for a shift towards examining *specific examples* of technology-enhanced workplace learning *implementations* aimed at *individual* learning processes within *distinct workplace contexts* to obtain a more nuanced understanding of the disparities in technology-enhanced workplace learning depending on the type of workplace. This approach allows us to shed light on the research question how access to digital learning opportunities is affected by the type of institution and the professional field.

B. Research Design

We conducted a rapid review [27] to evaluate the digital divide in workplace learning. Rapid reviews, which fall within the framework of Cochrane review methods [28, p. 5], are a streamlined approach to gathering evidence through synthesis and have a shorter turnaround time compared to traditional systematic reviews. In what follows, we explain the search and selection strategy that we derived from the objectives of this study – to describe the digital divide in workplace learning. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses [28][29] (PRISMA) approach was adopted to guide the screening process (see Figure 1).

The search strategy was as follows: We identified peer-reviewed journal publications published in the English language, and focused on technology-enhanced learning at the workplace. We used the Web of Science (WoS) online database to search for relevant publications, as this database matched best our search strategy and promised an efficient identification of relevant publications (contains peer-reviewed journal publications). Only publications published in 2020 or later were included. This is because we assume that the implementation of ALT in the workplace has undergone structural changes as a result of the COVID-19 pandemic. Review articles were excluded, as we are interested in institutional-level implementations of technology-enhanced learning.

Our search string (see also Table 1) refers to different synonyms of e-learning, and made reference to real-life ALT applications in a workplace setting. The search terms underwent further refinement and revised by an information specialist at the Brandenburg University of Applied Sciences. The final search string included restrictions (e.g., students at higher education institutions, pupils at schools, machine-

learning applications) for settings that do not classify as workplace learning.

Searches were conducted from February 16, 2024, to February 26, 2024, and yielded a total of 561 records (no duplicates). To account for the skewed distribution of publications on ALT towards the educational and health sectors, we conducted three separate searches for technology-enhanced learning (ibid.). These searches were conducted for educational institutions (N=130; 23% of records), for the health and medical sector (N=238; 42%) and for all other fields (N=193; 35%).

We recognize that this first step is merely an approximation, as we had not yet screened out records based on titles, keywords, and abstracts that may not be related to the use of ALT at the workplace. However, considering the high frequency of articles related to education and health and medical fields, and recognizing that most institutions in these fields are likely to be large and public sector-based, we believe that this approximation falls within the efficiency required by the chosen methodology (rapid reviews) while still retaining substantial validity for assessing the digital divide in workplace learning.

Table 1: Construction of the search string

Explanation	Components of the WoS search string		
<i>online learning or synonymous term in title referring to the use of ALT</i>	(((TI= ("digital" OR "virtual" OR "online" OR "hybrid" OR "remote" OR "blended" OR "distance" OR "web-based") AND TI= ("learning" OR "training" OR "course*") OR TI= ("e-learning" OR "elearning" OR "e-training" OR "entraining" OR "microlearning" OR "micro-learning" OR "mobile learning" OR "mobile-learning" OR "learning app"))		
	AND		
<i>concrete implementations...</i>	(AB= ("case stud*" OR "company case*" OR "field stud*" OR "field experiment*" OR "questionnaire*" OR "survey*") OR TI= ("case stud*" OR "company case*" OR "field stud*" OR "field experiment*" OR "questionnaire*" OR "survey*"))		
	AND		
<i>... at the workplace</i>	(AB= ("workplace*" OR "business*" OR "industry*" OR "industries" OR "enterprise*" OR "compan*" OR "public service*" OR "public sector*" OR "civil serv*" OR "corporat*" OR "professional*" OR "SME*" OR "governm*" OR "continuing education") OR TI= ("employee*" OR "worker*"))		
<i>exclude ALT applications aimed at students or pupils as well as machine learning applications</i>	NOT AB= ("student" OR "students" OR "pupil*" OR "machine learning" OR "deep learning" OR "reinforcement learning") NOT TI= ("student" OR "students" OR "pupil*" OR "machine learning" OR "deep learning" OR "reinforcement learning")		
	AND	AND	NOT
<i>relate to one of the three fields</i>	education	health and medical field	other fields
	AND	AND	NOT
	(TI= (teacher* OR faculty* OR lecturer*) OR AB= (teacher* OR faculty* OR lecturer*))	(AB= ("health*" OR "care" OR "medic*" OR "surg*" OR "radiol*" OR "dementia*" OR "clinic*" OR "nurse*") OR TI= ("health*" OR "care" OR "medic*" OR "surg*" OR "radiol*" OR "dementia*" OR "clinic*" OR "nurse*"))	(TI= (teacher* OR faculty* OR lecturer*) OR AB= (teacher* OR faculty* OR lecturer*)) NOT AB= ("health*" OR "care" OR "medic*" OR "surg*" OR "radiol*" OR "dementia*" OR "clinic*" OR "nurse*") NOT TI= ("health*" OR "care" OR "medic*" OR "surg*" OR "radiol*" OR "dementia*" OR "clinic*" OR "nurse*"))

The screening strategy for the 193 records resulting from the search for other sectors was as follows: Titles, keywords, and abstracts were screened for each record. Records that did not mention 'online' in connection with 'learning' (N=24), were not related to workplace learning or did not contain detailed information about a specific implementation at the workplace (N=81), excluding, e.g., studies focusing on organizational learning processes rather than individual learners' competency building, and studies that discuss abstract concepts or the interplay of general constructs in technology-enhanced workplace learning. Furthermore, we excluded studies without information about sector or professional field (N=27). This meant, e.g., that we exclude cross-sectional studies covering a large number of different institutions.

Furthermore, we identified additional review studies that have not been excluded in the initial WoS search routine (N=3). Similarly, we excluded further studies that refer to education (N=16) or to the health and medical field (N=10) that still ended up in the search results for "other sectors".

The remaining N=22 publications were included in full text screening. We excluded two additional studies because they were implemented and/or tested in a higher education context. Another was focused on knowledge management with MS Office and social media tools rather than with technology-enhanced learning. Moreover, we found two studies using the same ALT implementation example that were treated as duplicates and excluded one of them.

The 18 final full-text records underwent detailed analysis to gain systematic evidence on the digital divide in workplace learning. The screening was conducted with respect to the characteristics of the institution and the workplace, such as size, sector, and type of job.

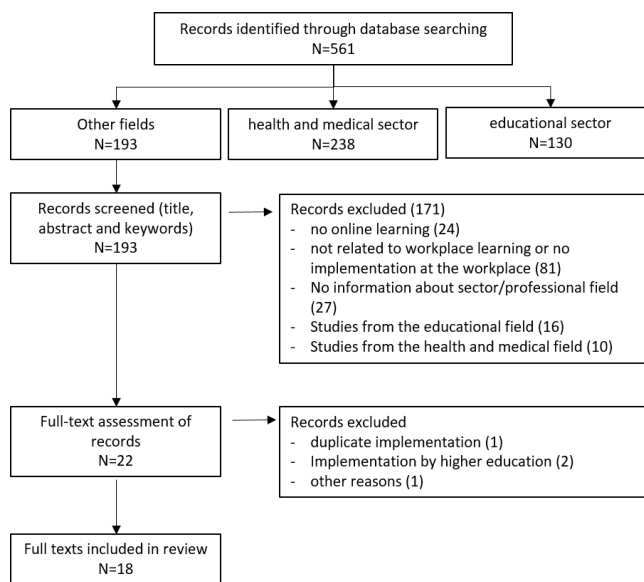


Figure 1: PRISMA-Chart

C. Results

Initially, the scarcity of studies on advanced learning technologies for workplace learning beyond higher education and healthcare is noteworthy. This scarcity suggests that – at least evidence-based and scientifically evaluated – implementation of ALT in the workplace is not yet that widespread, as we would expect given the generally acknowledged importance of reaping the benefits of ALT for workplace learning. Full-text screening of the 17 relevant studies identified yields the following picture (see also Table 2): The great majority of examples of ALT use at the workplace refers to large organizations or to cross-institutional implementations with participants from several institutions (e.g., engineers or agricultural workers employed in different companies or being self-employed). Our sample only contains one example at a medium-sized enterprise, and none at a small organization. Moreover, most applications are from technical sectors, such as energy, engineering, or automotive rather than from the service sector. Immersive virtual reality training (single or multi-player) is the most frequently found ALT, followed by mobile and micro-learning implementations. The picture becomes even clearer when we look at the occupational fields targeted by ALT in the records studied: it is mainly blue-collar workers who have access to ALT, especially VR-based immersive training, while service sector companies rather tend to use less technically sophisticated learning technologies, such as mobile and micro-learning. This discrepancy could be attributed to a generally lower level of digital literacy required to implement digitally supported learning in many service companies compared to technical, manufacturing companies where production processes have been increasingly automated with the help of digital technologies over the years. However, this hypothesis is challenged by the absence of companies from highly automated and digitized service sectors (e.g., finance, banking) in our literature review.

Table 2: Distribution of ALT-Use at the Workplace

Study	Institution	Sector	Profession	ALT	COLL
[31]	N/A	retail	diverse	MicroL	no
[32]	large	public	diverse	MobileL	yes
[33]	several	engineering	engineers	other	yes
[34]	large	business services	white-collar	other	yes
[35]	large	automotive	blue-collar	VR	no
[36]	large	public	white-collar	MicroL	no
[37]	several	IT	IT specialists	MicroL	no
[38]	large	public	both	other	no
[39]	medium	industrial services	blue-collar	VR	yes
[40]	N/A	food	N/A	other	no
[14]	large	energy	blue-collar	VR	no
[41]	N/A	energy	blue-collar	VR	no
[42]	large	steel	blue-collar	VR	no
[43]	several	electronics	blue-collar	VR	no
[44]	several	education	other	other	no
[45]	several	public	blue-collar	VR	yes
[46]	several	agriculture	blue-collar	MobileL	yes
[47]	large	chemical	diverse	other	N/A

Notes: ALT = advanced learning technologies, COL = collaborative learning, MicoL = micro-learning, MobileL = mobile learning, VR = Immersive virtual reality training

A third of the records analyzed cover ALT that fosters networked learning, i.e., collaboration between learners. Here, we cannot find differences in the use of ALT between white-collar and blue-collar professions.

D. Discussion

As a conclusion, our screening of the literature has revealed that there is a lack of ALT implementation at the workplace in other sectors, at least in terms of implementations that have been scientifically evaluated and the results have been published in peer-reviewed journal articles. Our results show that technology-enhanced learning opportunities are less frequent in smaller organizations, non-technical sectors (including the public sector) and for white-collar workers.

A major limitation of our analysis is publication bias. We may assume that the likelihood of writing an academic publication and publishing it in a peer-reviewed journal is higher in academic fields, such as health and education, which may partly explain the great number of results on the use of ALT for workplace learning we found.

Still despite these methodological limitations, our results indicate that there seems to be a digital divide in workplace learning, in particular along employer size and technological sector. Given that for example in Europe, almost two thirds of the employed workforce is working in small or medium-sized enterprises [48], and similarly, almost three quarters are employed in the service sector [49, p. 48], this poses a threat to workforce up- and reskilling and may severely hamper learning opportunities and individual development and growth for employees at such workplaces.

III. AI-BASED CONTENT CREATION AS A SOLUTION

A. The challenge of creating content for digital learning

To reduce the digital divide in workplace learning, it is essential to identify the obstacles preventing enterprises, particularly SMEs and service sector enterprises, from adopting digital learning technologies for the continuous up-skilling and re-skilling of their workforces. Previous studies have primarily focused on the lack of adequate technological infrastructure or general resistance to the introduction of e-learning for workplace learning [50].

A further explanation, which has not been fully investigated in previous studies of the adoption of advanced learning technologies, is that the burden on organizations of creating content specifically for an organization's needs may be a significant obstacle to the implementing digital learning in the workplace.

The creation of high-quality digital learning materials requires a combination of specialized skills in instructional design, multimedia production [51][52], and subject matter expertise [53]. Furthermore, the process of designing, developing and iterating digital content is time-consuming. It is estimated that between 40 and several hundred (!) hours may be required to develop one hour of e-learning content,

depending on the complexity of the material and the technologies used [54]-[56].

In summary, 'breathing life' into workplace ALT by creating high-quality learning content tailored to the organization's is a significant investment in human and financial resources. These non-technical barriers may slow down the adoption and implementation of digital learning solutions, limiting their potential benefits in improving workplace learning and development. A more detailed examination of the challenges of learning content creation also provides further insights into the propensity of organizations to adopt ALT for workplace learning, as discussed in our rapid literature review in Section II:

- Training content in companies varies significantly between the service and manufacturing sectors. A recent, cross-European study conducted by the OECD [57] found that the adoption of online delivery strongly depends on the content of training: Online delivery is most common for health, safety and security requirements as well as IT skills, but less so for technical, practical or job-specific skills such as machine or product training, sales training or customer handling, as well as soft skills such as communication, leadership, teamwork or conflict management. The creation of digital learning content for service sector topics may present greater challenges due to the need for interactive and scenario-based training that simulates real-life customer interactions and communication skills, which are inherently dynamic and context-specific. In contrast, manufacturing sector training may often involve more standardized and procedural content, such as safety protocols and technical skills, which are easier to codify and deliver as digital learning content. This could explain why service sector companies have been found to be less active in the provision of ALT at the workplace.
- SMEs are particularly reliant on informal learning [50][57]-[59], and make less use of classical training activities in classroom-like settings. Reasons for this preference may include limited resources for providing formal training to employees. Another explanation is that SMEs often offer jobs with a high task variety and excellent learning opportunities [59]. In addition, physical and social proximity is greater in smaller organizations, which provides particularly good conditions for informal learning [60] through feedback, trial and error, and observation of colleagues. However, the learning content conveyed by such informal learning activities is highly specific to the work process and activity concerned, and thus qualifies as highly contextualized material, characterized by above-average production times and costs. This may be an additional explanation for the lower propensity of SMEs to use digital learning approaches for their employees.

In conclusion, addressing the complexity and resource requirements for the creation of high quality, context-specific digital learning materials is crucial to fostering greater

adoption of advanced learning technologies in different organizational settings. By overcoming these barriers, SMEs and service sector companies, in particular, could benefit significantly, leading to a reduction in the digital divide in workplace learning.

B. AI-based creation of workplace learning content

Recently, the development and adoption of large language models (LLMs), such as OpenAI's ChatGPT, have attracted considerable interest for their ability to generate human-like conversational text content. These advances in generative AI can help automate the creation of high-quality, contextualized learning content [60]-[63]. There are promising applications for the automated generation of comprehensive learning content such as curricula [64][65], learning paths or course outlines, narrative educational elements [66], and interactive activities such as quizzes and reflection questions [67]. Significant potential is also attributed to the creation of personalized learning experiences [62][68], which tailor the pedagogical approach the specific abilities, interests, requirements and even learning styles of each student. Some of these approaches work based on a zero-shot basis, i.e., without the need to pre-train the AI model [69].

A number of commercial LLM-based content creation tools have emerged, enabling the creation of educational materials and comprehensive courses (for an overview of tools mainly aimed at school and academic use, see [70]. In addition to tools designed for education in schools and academia, there are more general-purpose AI-based course builders, such as Coursera's AI-based Course Builder [70], EdApp's AI Create [71], H5P's smart import [72], nolej.io [73], mindsmith.ai [74], and many others [75], which provide adequate functionality for designing digital learning for lifelong and workplace learning. The promise of these tools is to significantly reduce the time and cost needed of creating engaging, customized learning content [76].

However, AI-based automated content generation has primarily been used by educational institutions. Holmes and Littlejohn [77] note that, for this reason, AI in professional learning is primarily used to automate content creation in formal training courses with predefined content and outcomes, and is not yet widely used in informal workplace learning.

This paper outlines the significant potential of AI in creating digital learning content tailored to the specific needs of the workplace. As we will demonstrate, it is critical to capitalize on this potential, particularly in light of the observations made in Section II. These indicate that smaller organizations and service sector companies are lagging behind in the adoption of digital learning in the workplace. There are three main drivers for the adoption of digital learning in the workplace are as follows:

1. *Democratization through the reduced resource intensity of using AI support.* The use of AI to support the creation of digital learning content has the potential to democratize the design and delivery of digital learning activities. There is no need for subject matter experts or experienced staff to

have expertise in complex authoring software or to meet prerequisites in instructional design and pedagogical strategies. Furthermore, when sharing knowledge for AI-based content creation, language expression and formal correctness play only a minor role in the final quality of the material. This reduces the barriers for employees with limited language skills or little experience in formulating texts, enabling them to become AI-powered digital learning authors. AI-powered digital learning content creation can reduce production time and costs to a fraction of what they are today. AI-based automation also enables individuals and organizations to share their knowledge when time, financial resources, and e-learning skills are limited.

2. *Resource-efficient creation of company-specific learning content.* Significantly improved ability to leverage and frequently update customized, company-specific knowledge as often as needed with minimal effort. AI provides a vastly improved ability to leverage and frequently update customized, company-specific knowledge with minimal effort. Many AI course authoring tools allow uploading of a variety of formats, including text, video, and audio [66]. This flexibility means that existing corporate materials, such as product descriptions, technical descriptions, safety instructions, or anonymized customer complaint records, can be easily transformed into company-specific learning content that closely reflects organizational specifics and real-world work processes. Using such custom source material as a baseline produces learning content that is not only tailored to the specific needs of the prospective learners, but also improves the accuracy of the learning material because the AI is less prone to “hallucinations.”
3. *The power of generative AI tools to create human-like conversational content.* Another advantage of using generative AI tools to create digital learning content is their human-like conversational style. Generative AI's strengths in simulating human interactions [78] offer significant potential for creating high-quality materials such as interactive scenarios and digital role-plays, especially in soft skills and sales training. For example, AI-based role-plays could be developed using difficult customer scenarios based on common complaints or recorded audio from support calls.

In addition to these three main areas, AI exhibits considerable potential in the automated generation of learning content across a number of dimensions. These include multilingual learning units, which are becoming increasingly important given the international nature of many workforces. Content can also be tailored both didactically and contextually to different groups of learners (e.g., trainees, experienced learners, career changers) and different learning styles (e.g., experimental, visual). Adapting the instructional approach to each learner's individual learning style, progress, or skill level can be a valuable approach for workplace learning.

C. Discussion

In Section III.A, we examined the barriers to the creation of high-quality, context-specific digital learning materials for SMEs and service sector firms, highlighting that this may contribute to the digital divide in workplace learning, particularly with respect to the lower propensity to use ALT in smaller organizations and service sector firms. In Section III.B, we then explored how AI-based automated content generation can address these challenges by democratizing content creation, reducing the effort required to create and update customized learning content from unstructured sources, and leveraging human-like conversational styles to enhance interactive digital training for non-technical topics.

Our analysis suggests that AI has the potential to reduce the time and cost of producing digital learning content, while also supporting with the digitization of informal learning processes to some extent. AI's ability to transform unstructured materials, observations, and feedback, as well as process-specific content, into structured digital materials and maintain them at a relatively low cost and in a short time. This is particularly promising for small and medium-sized enterprises (SMEs), companies operating in highly dynamic environments, and companies in the service sector. This could narrow the size- and sector-dependent part of the digital divide in workplace learning identified in our rapid literature review in Section II.

Nevertheless, it is imperative to undertake a comprehensive assessment of the potential impediments to the narrowing of the digital divide in the context of workplace learning, with a particular focus on the burden of creating learning content. AI tools for automated content creation need to meet high standards: they should source custom material from flexible resources, generate human-like conversational output, adhere to strict data privacy standards when processing sensitive corporate information, be user-friendly for content creators, and be reasonably priced. Ideally, they should also use evidence-based instructional strategies relevant for effective digital (workplace) learning [24][79]. Some researchers call for “pedagogical intelligence” to work hand-in-hand with artificial intelligence in education, criticizing the lack of pedagogical foundations guiding current AI research in schools [80].

After testing some of the existing AI tools mentioned above, it is clear that no current tool meets all these requirements (for privacy concerns, see [69]). Many tools are promising in their functionalities. Even if adequate tools were readily available and accessible, quality concerns could still be an issue. A common solution is to use a human-in-the-loop approach [76][81], where human experts manually review learning content at critical processing stages.

IV. CONCLUSION

In summary, the digital workplace divide remains a significant issue, particularly for smaller organizations with limited resources and those requiring highly specific, non-technical training content. The high demands on time, money, and human capital to produce company-specific learning

content have been identified as a major cause of this divide. Automating content creation using generative AI offers a promising solution to narrow this gap.

It is of the utmost importance that policymakers, society, and industry work in collaboration to prevent the potential exacerbation of existing inequalities through the creation of an "AI divide" in the workplace. Consequently, it is imperative that AI-based learning content creation evolve into an inclusive technology. This would enable a broader range of workers to act as subject matter experts and share knowledge through self-created digital learning materials, thereby promoting widespread workplace learning opportunities for all workers and narrowing the digital divide in workplace learning.

REFERENCES

- [1] K. Frosch, F. Lindauer, and N. Zuidhof, "Exploring the Digital Divide in Workplace Learning: a Rapid Review," presented at the ICDS 2024, The Eighteenth International Conference on Digital Society, May 2024, pp. 1–6. [Online]. Available: https://www.thinkmind.org/library/ICDS/ICDS_2024/icds_2024_1_10_18001.html (last accessed November 6, 2024).
- [2] K. Smet, I. Grosemans, N. D. Cuyper, and E. Kyndt, "Outcomes of Informal Work-Related Learning Behaviours: A Systematic Literature Review," *Scandinavian Journal of Work and Organizational Psychology*, vol. 7, Art. no. 1, Feb. 2022, doi: 10.16993/sjwop.151.
- [3] A. Clardy, "70-20-10 and the Dominance of Informal Learning: A Fact in Search of Evidence," *Human Resource Development Review*, vol. 17, no. 2, pp. 153–178, Jun. 2018, doi: 10.1177/1534484318759399.
- [4] V. Alevan, C. R. Beal, and A. C. Graesser, "Introduction to the Special Issue on Advanced Learning Technologies," *J. Educ. Psychol.*, vol. 105, no. 4, pp. 929–931, Nov. 2013, doi: 10.1037/a0034155.
- [5] T. Ley, "Knowledge Structures for Integrating Working and Learning: A Reflection on a Decade of Learning Technology Research for Workplace Learning," *British Journal of Educational Technology*, vol. 51, no. 2, pp. 331–346, 2020, doi: 10.1111/bjet.12835.
- [6] M. Werkle, M. Schmidt, D. Dikke, and S. Schwantzer, "Case Study 4: Technology enhanced Workplace Learning," *Responsive open learning environments: Outcomes of research from the ROLE project*, pp. 159–184, 2015.
- [7] K. Frosch, "Scan to Learn: a Lightweight Approach for Informal Mobile Micro-Learning at the Workplace," presented at the eLmL 2023, The Fifteenth International Conference on Mobile, Hybrid, and On-line Learning, Apr. 2023, pp. 53–61. [Online]. Available: https://www.thinkmind.org/index.php?view=article&articleid=elml_2023_2_80_50039 (last accessed November 6, 2024).
- [8] P. Vassilakopoulou and E. Hustad, "Bridging Digital Divides: a Literature Review and Research Agenda for Information Systems Research," *Inf Syst Front*, vol. 25, no. 3, pp. 955–969, Jun. 2023, doi: 10.1007/s10796-020-10096-3.
- [9] S. Lythreatis, S. K. Singh, and A.-N. El-Kassar, "The Digital Divide: a Review and Future Research Agenda," *Technological Forecasting and Social Change*, vol. 175, p. 121359, 2022.
- [10] A. Cheshmehzangi, T. Zou, Z. Su, and T. Tang, "The growing Digital Divide in Education among primary and secondary Children during the COVID-19 Pandemic: An Overview of Social Exclusion and Education Equality Issues," *Journal of Human Behavior in the Social Environment*, vol. 33, no. 3, pp. 434–449, Apr. 2023, doi: 10.1080/10911359.2022.2062515.
- [11] A. Mathrani, T. Sarvesh, and R. Umer, "Digital Divide Framework: Online Learning in Developing Countries During the COVID-19 Lockdown," *Globalisation, Societies and Education*, vol. 20, no. 5, pp. 625–640, Oct. 2022, doi: 10.1080/14767724.2021.1981253.
- [12] A. D. Ritzhaupt, L. Cheng, W. Luo, and T. N. Hohlfeld, "The Digital Divide in Formal Educational Settings: the Past, Present, and Future Relevance," in *Handbook of Research in Educational Communications and Technology*, M. J. Bishop, E. Boling, J. Elen, and V. Svihla, Eds., Cham: Springer International Publishing, 2020, pp. 483–504. doi: 10.1007/978-3-030-36119-8_23.
- [13] E. Kormos and K. Wisdom, "Rural Schools and the Digital Divide: Technology in the Learning Experience," *Theory & Practice in Rural Education*, vol. 11, no. 1, pp. 25–39, 2021.
- [14] P. Gorski, "Education equity and the digital divide," *AACE Review (Formerly AACE Journal)*, vol. 13, no. 1, pp. 3–45, 2005.
- [15] P. Tynjälä, "Perspectives into Learning at the Workplace," *Educational Research Review*, vol. 3, no. 2, pp. 130–154, Jan. 2008, doi: 10.1016/j.edurev.2007.12.001.
- [16] M. Gaeta, V. Loia, F. Orciuoli, and P. Ritrovato, "S-WOLF: Semantic Workplace Learning Framework," *IEEE Trans. Syst. Man Cybern. -Syst.*, vol. 45, no. 1, pp. 56–72, Jan. 2015, doi: 10.1109/TSMC.2014.2334551.
- [17] L. M. English and A. Carlsen, "Lifelong Learning and the Sustainable Development Goals (SDGs): Probing the Implications and the Effects," *Int Rev Educ*, vol. 65, no. 2, pp. 205–211, Apr. 2019, doi: 10.1007/s11159-019-09773-6.
- [18] A. Granić, "Educational Technology Adoption: A systematic review," *Educ Inf Technol*, vol. 27, no. 7, pp. 9725–9744, Aug. 2022, doi: 10.1007/s10639-022-10951-7.
- [19] K. Frosch and N. Zuidhof, "Using Mobile Technologies for Situating Bite-Sized Learning at the Workplace", *International Journal On Advances in Intelligent Systems*, vol. 16, no. 3 and 4, pp. 89–101. [Online]. Available: https://www.thinkmind.org/library/IntSys/IntSys_v16_n34_2023/intsys_v16_n34_2023_7.html (last accessed November 6, 2024).
- [20] M. N. Giannakos, P. Mikalef, and I. O. Pappas, "Systematic Literature Review of E-Learning Capabilities to Enhance Organizational Learning," *Inf Syst Front*, vol. 24, no. 2, pp. 619–635, Apr. 2022, doi: 10.1007/s10796-020-10097-2.
- [21] W. Yu, J. He, and Y. Gong, "A Systematic Review of Information Technology in Workplace Learning,"

- Psychology In The Schools*, Mar. 2023, doi: 10.1002/pits.22901.
- [22] M. M. Serema, D. S. P. Shihomeka, and R. K. Shalyefu, "Adoption and Utilisation of Workplace E-Learning Practices in the Public Sector Organisations: A Systematic Review," *Journal of Learning for Development*, vol. 10, no. 3, Art. no. 3, Nov. 2023, doi: 10.56059/jl4d.v10i3.944.
- [23] M. Hogeфорster and M. Wildt, "Leveraging Digital Learning and Work-Based Learning to enhance Employee Skills in Small and Medium Enterprises," presented at the 13th International Scientific Conference „Business and Management 2023“, Vilnius Gediminas Technical University, Lithuania, May 2023. doi: 10.3846/bm.2023.1001.
- [24] B. Kaizer, C. da Silva, T. Zerbin, and A. Paiva, "E-learning Training in Work Corporations: A Review on Instructional Planning," *European Journal of Training and Development*, vol. 44, no. 8–9, pp. 761–781, Oct. 2020, doi: 10.1108/EJTD-03-2020-0042.
- [25] R. Alfaleh, L. East, Z. Smith, and S.-Y. Wang, "Nurses? Perspectives, Attitudes and Experiences related to E-Learning: A Systematic Review," *Nurse Education Today*, vol. 125, Jun. 2023, doi: 10.1016/j.nedt.2023.105800.
- [26] S. Wu, "Unpacking Themes of Integrating Telecollaboration in Language Teacher Education: A Systematic Review of 36 Studies from 2009 to 2019," *Computer Assisted Language Learning*, vol. 36, no. 7, pp. 1265–1287, Sep. 2023, doi: 10.1080/09588221.2021.1976800.
- [27] B. Smela *et al.*, "Rapid Literature Review: Definition and Methodology," *Journal of Market Access & Health Policy*, vol. 11, no. 1, Art. no. 1, Jan. 2023, doi: 10.1080/20016689.2023.2241234.
- [28] N. Zuidhof, S. Allouch, O. Peters, and P.-P. Verbeek, "Defining Smart Glasses: A Rapid Review of State-of-the-Art Perspectives and Future Challenges from a Social Sciences' Perspective," *Augmented Human Research*, vol. 6, Dec. 2021, doi: 10.1007/s41133-021-00053-3.
- [29] D. Moher *et al.*, "Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 statement," *Syst Rev*, vol. 4, no. 1, p. 1, Jan. 2015, doi: 10.1186/2046-4053-4-1.
- [30] L. Shamseer *et al.*, "Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015: Elaboration and Explanation," *Bmj*, vol. 349, 2015. [Online]. Available: <https://www.bmj.com/content/349/bmj.G7647.abstract> (last accessed November 6, 2024).
- [31] J. Karlsen, E. Balsvik, and M. Ronnevik, "A Study of Employees' Utilization of Microlearning Platforms in Organizations," *Learning Organization*, vol. 30, no. 6, pp. 760–776, Nov. 2023, doi: 10.1108/TLO-07-2022-0080.
- [32] A. Dolowitz, J. Collier, A. Hayes, and C. Kumsal, "Iterative Design and Integration of a Microlearning Mobile App for Performance Improvement and Support for NATO Employees," *Techtrends* vol. 67, no. 1, pp. 143–149, Jan. 2023, doi: 10.1007/s11528-022-00781-2.
- [33] T. Li *et al.*, "Learning MBSE Online: A Tale of Two Professional Cohorts," *Systems*, vol. 11, no. 5, Apr. 2023, doi: 10.3390/systems11050224.
- [34] G. Angafor, I. Yevseyeva, and L. Maglaras, "Scenario-based Incident Response Training: Lessons Learnt from Conducting an Experiential Learning Virtual Incident Response Tabletop Exercise," *Information and Computer Security*, vol. 31, no. 4, pp. 404–426, Oct. 2023, doi: 10.1108/ICS-05-2022-0085.
- [35] J. Monteiro, D. Torres, A. Ramos, and C. Pimentel, "The use of Virtual Reality as E-Training Tool for dies' Changeover in Stamping Presses: a Case Study on Automotive Industry," *International Journal of Lean Six Sigma*, Dec. 2023, doi: 10.1108/IJLSS-02-2023-0041.
- [36] T. Beste, "Knowledge Transfer in a Project-Based Organization Through Microlearning on Cost-Efficiency," *Journal of Applied Behavioral Science*, vol. 59, no. 2, pp. 288–313, Jun. 2023, doi: 10.1177/002188632111033096.
- [37] R. Gerbaudo, R. Gaspar, and R. Lins, "Novel Online Video Model for Learning Information Technology Based on Micro Learning and Multimedia Micro Content," *Education and Information Technologies*, vol. 26, no. 5, pp. 5637–5665, Sep. 2021, doi: 10.1007/s10639-021-10537-9.
- [38] I. Casebourne, "Left to their own Devices: An Exploration of Context in Seamless Work-Related Mobile Learning," *British Journal of educational Technology*, Jan. 2024, doi: 10.1111/bjet.13410.
- [39] S. Pedram, S. Palmisano, S. Mielle, M. Farrelly, and P. Perez, "Influence of Age and Industry Experience on Learning Experiences and Outcomes in Virtual Reality Mines Rescue Training," *Frontiers in virtual Reality*, vol. 3, Oct. 2022, doi: 10.3389/frvir.2022.941225.
- [40] P. Sureephong, W. Dahlan, S. Chernbumroong, and Y. Tongpaeng, "The Effect of Non-Monetary Rewards on Employee Performance in Massive Open Online Courses," *International Journal of emerging Technologies in Learning*, vol. 15, no. 1, pp. 88–102, 2020, doi: 10.3991/ijet.v15i01.11470.
- [41] I. Bernal *et al.*, "An Immersive Virtual Reality Training Game for Power Substations Evaluated in Terms of Usability and Engagement," *Applied Sciences-Basel*, vol. 12, no. 2, Jan. 2022, doi: 10.3390/app12020711.
- [42] K. Dhalmahapatra, J. Maiti, and O. Krishna, "Assessment of Virtual Reality based Safety Training Simulator for electric overhead Crane Operations," *Safety Science*, vol. 139, Jul. 2021, doi: 10.1016/j.ssci.2021.105241.
- [43] Y. Kang, H. Song, H. Yoon, and Y. Cho, "The Effect of Virtual Reality Media Characteristics on Flow and Learning Transfer in Job Training: The Moderating Effect of Presence," *Journal of Computer assisted Learning*, vol. 38, no. 6, pp. 1674–1685, Dec. 2022, doi: 10.1111/jcal.12702.
- [44] B. Bruijns *et al.*, "Change in pre- and in-service early Childhood Educators' Knowledge, Self-Efficacy, and Intentions Following an E-Learning Course in Physical Activity and Sedentary Behaviour: a Pilot Study," *BMC Public Health*, vol. 22, no. 1, Feb. 2022, doi: 10.1186/s12889-022-12591-5.

- [45] D. Narciso, M. Melo, S. Rodrigues, J. P. Cunha, J. Vasconcelos-Raposo, and M. Bessa, "Using Heart Rate Variability for comparing the Effectiveness of Virtual vs Real Training Environments for Firefighters," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 7, pp. 3238–3250, Jul. 2023, doi: 10.1109/TVCG.2022.3156734.
- [46] N. Zamani, F. Kazemi, and E. Masoomi, "Determinants of entrepreneurial Knowledge and Information sharing in Professional Virtual Learning Communities created using Mobile Messaging Apps," *Journal of Global Entrepreneurship Research*, vol. 11, no. 1, pp. 113–127, Dec. 2021, doi: 10.1007/s40497-021-00275-0.
- [47] R. Aragao, C. Pereira-Guzzo, and P. Figueiredo, "Impacts of an E-Learning System on the Occurrence of Work Accidents in a Chemical Industry Company," *International Journal of Knowledge Management Studies*, vol. 11, no. 4, pp. 325–343, 2020.
- [48] European Commission. Joint Research Centre. and European Commission. Directorate General for Internal Market, Industry, Entrepreneurship and SMEs., *Annual report on European SMEs 2022/2023: SME performance review 2022/2023*. LU: Publications Office, 2023. [Online]. Available: <https://data.europa.eu/doi/10.2760/028705> (last accessed November 6, 2024).
- [49] EUROSTAT, "Key Figures on Europe 2023 Edition," 2023. [Online]. Available: <https://ec.europa.eu/eurostat/web/products-key-figures/wks-ei-23-001> (last accessed November 6, 2024).
- [50] W. Admiraal and D. Lockhorst, "E-Learning in Small and Medium-Sized Enterprises Across Europe: Attitudes Towards Technology, Learning and Training," *International Small Business Journal*, vol. 27, no. 6, pp. 743–767, Dec. 2009, doi: 10.1177/0266242609344244.
- [51] A. D. Ritzhaupt and F. Martin, "Development and Validation of the Educational Technologist Multimedia Competency Survey," *Education Tech Research Dev*, vol. 62, no. 1, pp. 13–33, Feb. 2014, doi: 10.1007/s11423-013-9325-2.
- [52] J. D. Klein and W. Q. Kelly, "Competencies for instructional Designers: A View from Employers," *Performance Improvement Quarterly*, vol. 31, no. 3, pp. 225–247, 2018, doi: 10.1002/piq.21257.
- [53] S. Palmer, "Instructional Designers' Insights Regarding Factors that are important yet Lacking in Subject Matter Experts", Idaho State University, 2023. [Online]. Available: <https://etd.iri.isu.edu/ViewSpecimen.aspx?id=2264> (last accessed November 6, 2024).
- [54] B. Chapman, "How long does it take to create Learning?," 2010. [Online]. Available: [https://www.cedma-europe.org/newsletter%20articles/misc/How%20long%20does%20it%20take%20to%20develop%20training%20by%20Brian%20Chapman%20\(Sep%2010\).pdf](https://www.cedma-europe.org/newsletter%20articles/misc/How%20long%20does%20it%20take%20to%20develop%20training%20by%20Brian%20Chapman%20(Sep%2010).pdf) (last accessed November 6, 2024).
- [55] K. M. Kapp and R. A. Defelice, "Time to develop one hour of Training By Karl M. Kapp, Robyn A. Defelice." [Online]. Available: <https://www.td.org/newsletters/learning-circuits/time-to-develop-one-hour-of-training-2009> (last accessed November 6, 2024).
- [56] S. V. Naik and K. Laxman, "A Study on the Design/Development time of E-Learning Projects in New Zealand Sweety Viral Naik 1, Kumar Laxman." *J. adv. humanit. soc. sci.*, vol. 3, no. 1, Feb. 2017, doi: 10.20474/jahss-3.1.1.
- [57] OECD, *Training in Enterprises: New Evidence from 100 Case Studies*. Paris: Organisation for Economic Co-operation and Development, 2021. [Online]. Available: https://www.oecd-ilibrary.org/employment/training-in-enterprises_7d63d210-en (last accessed: November 6, 2024).
- [58] M. S. Cardon and S. D. Valentin, "16 Training and Development in Small and Medium Enterprises," *The Cambridge handbook of workplace training and employee development*, p. 363, 2017.
- [59] A. Coetzer, H. Kock, and A. Wallo, "Distinctive Characteristics of Small Businesses as Sites for Informal Learning," *Human Resource Development Review*, vol. 16, no. 2, pp. 111–134, Jun. 2017, doi: 10.1177/1534484317704291.
- [60] D. Baidoo-Anu and L. Owusu Ansah, "Education in the Era of Generative Artificial Intelligence (AI): Understanding the potential Benefits of ChatGPT in Promoting Teaching and Learning." Rochester, NY, Jan. 25, 2023. doi: 10.2139/ssrn.4337484.
- [61] T. K. F. Chiu, "The impact of Generative AI (GenAI) on Practices, Policies and Research Direction in Education: a Case of ChatGPT and Midjourney," *Interactive Learning Environments*, vol. 0, no. 0, pp. 1–17, 2023, doi: 10.1080/10494820.2023.2253861.
- [62] S. Grassini, "Shaping the Future of Education: exploring the Potential and Consequences of AI and ChatGPT in Educational Settings," *Education Sciences*, vol. 13, no. 7, p. 692, Jul. 2023, doi: 10.3390/educsci13070692.
- [63] D. T. T. Mai, C. V. Da, and N. V. Hanh, "The Use of ChatGPT in Teaching and Learning: a Systematic Review through SWOT Analysis Approach," *Front. Educ.*, vol. 9, Feb. 2024, doi: 10.3389/educ.2024.1328769.
- [64] A. Padovano and M. Cardamone, "Towards Human-AI Collaboration in the Competency-Based Curriculum Development Process: the Case of Industrial Engineering and Management Education," *Computers and Education: Artificial Intelligence*, vol. 7, p. 100256, Dec. 2024, doi: 10.1016/j.caeai.2024.100256.
- [65] G. van den Berg and E. du Plessis, "ChatGPT and Generative AI: Possibilities for its Contribution to Lesson Planning, Critical Thinking and Openness in Teacher Education," *Education Sciences*, vol. 13, no. 10, Art. no. 10, Oct. 2023, doi: 10.3390/educsci13100998.
- [66] C. Diwan, S. Srinivasa, G. Suri, S. Agarwal, and P. Ram, "AI-based Learning Content Generation and Learning Pathway Augmentation to increase Learner Engagement," *Computers and Education: Artificial Intelligence*, vol. 4, p. 100110, Jan. 2023, doi: 10.1016/j.caeai.2022.100110.

- [67] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A Systematic Review of Automatic Question Generation for Educational Purposes," *International Journal Artificial Intell Education*, vol. 30, no. 1, pp. 121–204, Mar. 2020, doi: 10.1007/s40593-019-00186-y.
- [68] I. Pesovski, R. Santos, R. Henriques, and V. Trajkovik, "Generative AI for customizable Learning Experiences," *Sustainability*, vol. 16, no. 7, p. 3034, Apr. 2024, doi: 10.3390/sul16073034.
- [69] L. Yan *et al.*, "Practical and ethical Challenges of Large Language Models in Education: A Systematic Scoping Review," *British Journal of Educational Technology*, vol. 55, no. 1, pp. 90–112, 2024, doi: 10.1111/bjet.13370.
- [70] Coursera, "Coursera for Business - Experience Custom Course Creation at Scale." [Online]. Available: <https://www.coursera.org/business/course-builder> (last accessed: November 6, 2024).
- [71] EdApp, "https://support.edapp.com/creating-courses-with-ai-create-beta," Create with AI Create. [Online]. Available: <https://support.edapp.com/creating-courses-with-ai-create-beta> (last accessed November 6, 2024).
- [72] H5P, "Introducing Smart Import." [Online]. Available: <https://campaigns.h5p.com/h5p-smart-import/> (last accessed November 6, 2024).
- [73] Nolej, "https://nolej.io/," NOLEJ - Level up your campus with AI! [Online]. Available: <https://nolej.io/> (last accessed November 6, 2024).
- [74] Mindsmith, "Mindsmith - accelerate your eLearning Development with Generative AI." [Online]. Available: <https://www.mindsmith.ai/> (last accessed November 6, 2024).
- [75] A. Burton, "7 Best AI-Powered Course Builders in 2024." [Online]. Available: usefulai.com/tools/ai-course-builders (last accessed November 6, 2024).
- [76] D. Leiker, S. Finnigan, A. R. Gyllen, and M. Cukurova, "Prototyping the Use of Large Language Models (LLMs) for Adult Learning Content Creation at Scale." arXiv, Jun. 02, 2023. doi: 10.48550/arXiv.2306.01815.
- [77] W. Holmes and A. Littlejohn, "Artificial Intelligence for Professional Learning," in *Handbook of Artificial Intelligence at Work*, Edward Elgar Publishing, 2024, pp. 191–211.
- [78] Z. Adiguzel and F. Cakir, "Investigation of the Effects of Learning and Performance Goal Orientation on Team Creative Efficacy and New Service Development Performance Within the Organization," *Revista de Estudios Empresariales-Segunda Epoca*, no. 2, pp. 230–250, 2022, doi: 10.17561/ree.n2.2022.6442.
- [79] R. E. Mayer, "Thirty Years of Research on Online Learning," *Applied Cognitive Psychology*, vol. 33, no. 2, pp. 152–159, 2019, doi: 10.1002/acp.3482.
- [80] B. Díaz and M. Nussbaum, "Artificial Intelligence for Teaching and Learning in Schools: The Need for pedagogical Intelligence," *Computers & Education*, vol. 217, p. 105071, Aug. 2024, doi: 10.1016/j.compedu.2024.105071.
- [81] T. Weinert, M. Billert, M. T. de Gafenco, A. Janson, and J. M. Leimeister, "Designing a Co-creation System for the Development of Work-process-related Learning Material in Manufacturing," *Comput Supported Coop Work*, vol. 32, no. 1, pp. 5–53, Mar. 2023, doi: 10.1007/s10606-021-09420-5.

Developing a Sign Language Writing System: Focus on Necessity and Sign Language-Specific Features

Nobuko Kato, Yuito Nameta, Megumi Shimomori,
Sumihiro Kawano, Yuhki Shiraishi
Faculty of Industrial Technology
Tsukuba University of Technology
Tsukuba, Japan
e-mail: {nobuko, a191009, a171013, kawano,
yuhkis}@a.tsukuba-tech.ac.jp

Akihisa Shitara
Graduate School of Library, Information and Media Studies
University of Tsukuba
Tsukuba, Japan
e-mail: theta-akihisa@digitalnature.slis.tsukuba.ac.jp

Abstract—Achieving universal access in professional settings necessitates the development of computer-assisted sign language writing support system, considering the perceptual characteristics of the deaf and hard of hearing individuals. This study explores sign language-specific features to elucidate the requirements for a sign language writing support system. Analysis of news sentences expressed in sign language reveals the prevalence of distinct expressions like topicalized and wh-cleft sentences. We explore a writing system that incorporates these features and conduct experiments involving transcribing sign language movies. We first examine whether it is necessary to write sign language when learning in specialized contexts, thus identifying the key features of sign language sentences that need to be written effectively and clarify the functions required for the system based on actual writing experiments.

Keywords—deaf and hard of hearing; sign language; visual language; sign writing; communication support.

I. INTRODUCTION

The purpose of this study was to develop a computer-based sign language writing support system tailored to the perceptual characteristics of deaf and hard-of-hearing (DHH) individuals, considering the visual and spatial nature of sign language and the unique characteristics of signed sentences. This study complements the previous work [1] by discussing the need for such a system and examining the issues involved in developing a computer-based writing support system for sign language writing.

The enrollment of individuals with disabilities in higher education institutions and the emphasis on lifelong learning are increasing, necessitating expanded learning opportunities tailored to individual disabilities. In specialized educational settings such as higher education, it is necessary to ensure that effective information and communication methods align with the unique characteristics of each disability.

Various services are employed to facilitate communication among DHH individuals in higher education institutions, including real-time captioning by transcriptionists, automatic speech recognition (ASR), sign language interpretation, and notetaking. ASR technology is increasingly being explored to automatically generate caption text for DHH users [2]. However, it is crucial to recognize that DHH individuals are bicultural and have the right to be

educated in their native sign language [3]. Quality education delivered in national sign languages and written languages is a key factor in the education of deaf children and adult learners [4].

While some countries use sign language with word orders that mirror spoken language, they can pose comprehension challenges for deaf individuals [5]. Research on sign language interpretation in universities has indicated that deaf students must receive information using the correct sign language structure [6]. Studies on sign language interpretation in universities have highlighted the significance of instructors' clear use of sign language, as perceived by deaf students [7].

In other words, it is considered important for quality education that direct instructors and sign language interpreters use the sign language accurately and that students receive information using the correct sign language structure. Consequently, there is an anticipated increase in opportunities for specialized content learning facilitated by interpreters or direct sign language instruction in various countries.

Writing has also been considered an important process in higher education. However, writing presents a significant challenge in sign language learning. Existing writing systems for spoken language (Figure 1d) are ill-suited for sign language, which is a distinct language. Unlike hearing individuals, who can write while listening (Figure 1a), deaf individuals must write while simultaneously watching sign language (Figure 1b).

Therefore, the development of a computer-based support system for writing sign language is essential for streamlining the writing process and allocating more time to the comprehension of specialized content. To achieve this, it is imperative to delineate the functions that such a system should encompass, based on sign language characteristics.

This study aims to address the following research questions:

RQ1: Is a new system for writing sign language necessary in professional and learning situations?

RQ2: What are the sign language-specific features crucial for writing specialized sign language content?

RQ3: How can sentences be written while preserving sign language-specific expressions?

RQ4: What challenges are faced by the proposed sign language writing method in developing a computer-based support system for writing sign language?

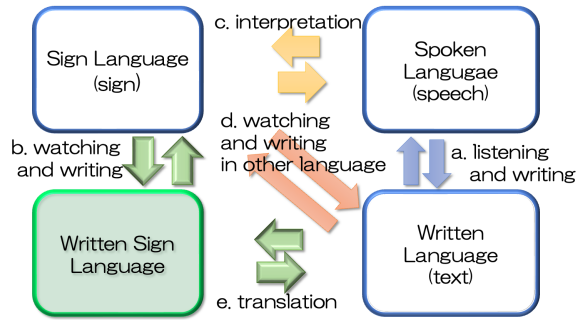


Figure 1. The relationship between spoken and written language.

The remainder of this paper is structured as follows: Section II provides insights into sign languages and relevant prior research. Section III describes the verifications to assess the necessity of a sign language writing system. Section IV outlines the characteristics of signed sentences and presents the proposed method based on these characteristics. Section V elaborates the experimental methodology and results, and Section VI discusses the findings based on the experimental results. Finally, Section VII summarizes this study.

II. SIGN LANGUAGE NOTATION METHODS

After discussing the differences between signed and spoken languages from the perspective of writing signs, this chapter reviews related previous studies.

A. Sign Language

Sign language serves as a visual language used by the deaf community, where linguistic information is communicated not only through hand shapes and movements but also through nonmanual markers (NMMs) such as facial expressions, gaze, and head movements [5].

Unlike spoken languages such as English, which are linear and rely on speech, sign languages are intricate and employ hand gestures, facial expressions, body movements, and spatial elements [5]. Thus, devising a writing system for sign languages demands innovative approaches that are distinct from those used for spoken languages.

B. Related Work

Efforts to transcribe sign language into writing have adopted two main approaches: iconographic and alphabetic (using letters from existing spoken languages) [8].

Iconographic methods entail symbolizing hand actions and describing words and sentences, offering the advantage of representing novel words and actions, but often result in a high number of descriptions per word, primarily suited for analysis [9][10]. Notational systems such as Si5S and ASLwrite prioritize writing, but use specialized fonts for sign language, which makes it difficult for learners to correlate these systems with existing spoken language texts.

ASL-gloss, another method, employs characters from existing spoken languages using English words as labels to describe American Sign Language (ASL). This system follows the ASL word order and grammatical rules, with glosses used to teach sign language and grammar [11]. Few

studies have examined the use of ASL-gloss in actual educational settings, and examples that have examined the use of ASL-gloss as a potential method for improving reading and comprehension skills in people with severe hearing loss have not supported it as an effective method for improving comprehension [12].

One example of using Japanese as a label is when it is used as an intermediate language for machine translation between Japanese and signed Japanese [13].

In university settings, where comprehension hinges on understanding key spoken words, it is crucial for deaf students to receive information that is semantically and syntactically correct in the sign language structure [6].

Therefore, our study adopts existing characters to describe terms, and explores a method for diagrammatically representing the structure of sign languages to address these challenges.

III. VERIFICATION IN A LEARNING ENVIRONMENT

DHH individuals who use sign language have been learning to use both sign language and Japanese since an early age, using Japanese texts. This raises the question of whether it is necessary to learn and write using sign language in professional and learning situations.

To address RQ1, verification was conducted using two types of sign language videos—Expressions 1 and 2—to assess the necessity of a sign language writing system in learning environments.

Expression 1: Signed words were arranged according to Japanese word order, along with Japanese mouthing and fingerspelling, and the sign language expressions for each word were taken from a sign language dictionary designed for learning.

Expression 2: Spatial and visual expressions were employed using Japanese Sign Language (JSL) grammar.

Ethical approval was obtained from the Research Ethics Review of Tsukuba University of Technology, where the experiments were conducted.

A. Verification 1

The video used in Verification 1 explained spatial geometry in mathematics and focused on explaining specialized content in sign language without using any materials or whiteboards, relying solely on sign language for the explanation. A deaf individual proficient in JSL created videos for Expressions 1 and 2 after fully understanding the content. The duration of the videos was 78.9 seconds for Expression 1 and 69.0 seconds for Expression 2.

Eight DHH students who regularly use sign language participated in the experiment. They watched the videos either in the order of Expression 1, followed by Expression 2, or in reverse order.

Immediately after watching each video, participants rated its clarity on a 7-point scale (1: very unclear, 4: neutral, 7: very clear) and attempted to explain the content of the video

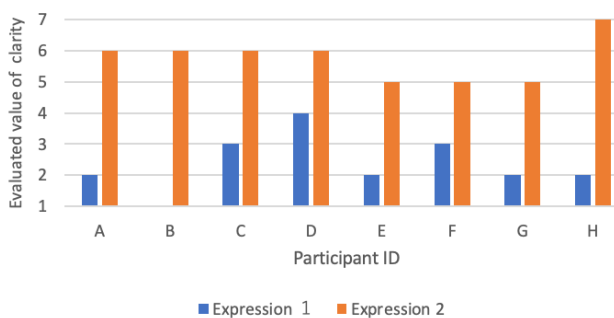


Figure 2. Results of the questionnaire on the clarity of Verification 1 (1: very unclear, 4: neutral, 7: very clear).

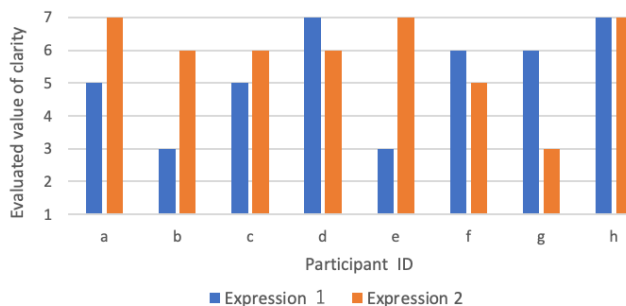


Figure 3. Results of the questionnaire on the clarity of Verification 2 (1: very unclear, 4: neutral, 7: very clear).

in their own words. Additionally, after viewing both videos, they answered a follow-up survey on each expression.

As a result of the evaluation, all eight participants rated Expression 2 as easier to understand than Expression 1 (Figure 2). The average clarity ratings were 2.4 for Expression 1 and 5.8 for Expression 2. A paired t-test revealed a significant difference ($p < 0.01$).

Additionally, in the follow-up survey, the following comment was made:

- In Expression 1, it was difficult to read Japanese (through mouthing or fingerspelling, etc.).
- In Expression 2, it was easier to form a visual image.

In the explanations provided by the experiment participants regarding the video content, it was observed that

- In Expression 1, they understood the mathematical terms (with examples in which they reproduced the Japanese directly from their notes).
- In Expression 2, providing a detailed explanation was more difficult.

The average accuracy of the reproduced sentences was 55% for Expression 1 and 52% for Expression 2, with no significant difference between them. In both cases, the reproduction rate was low.

In other words, although Expression 2, which used the grammar of JSL, was rated as easier to understand, accurately reproducing the content was difficult. It was confirmed that a method for writing sign language is necessary to fully comprehend and reproduce the content.

B. Verification 2

In Verification 1, we found differences in the level of mathematical knowledge among students. To eliminate the influence of prior knowledge, the content of the video was set to “Denseness of real numbers,” a topic that was unknown to all participants.

Participants were eight DHH students who regularly used sign language. Participants watched Expression 1 and responded to a comprehension test and questionnaire, then watched Expression 2 and responded to a comprehension test or questionnaire, or vice versa.

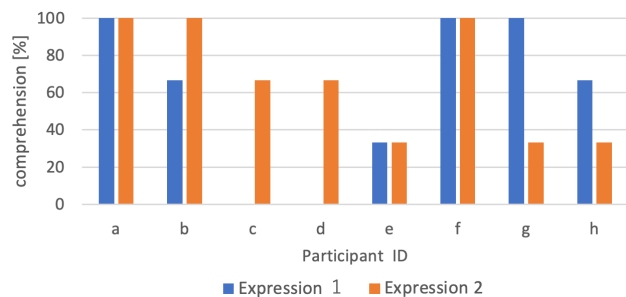


Figure 4. Comprehension in Verification 2.

TABLE I. RESULTS OF QUESTIONNAIRE AND COMPREHENSION TEST IN VERIFICATION 2.

	Expression 1	Expression 2
Average of questionnaire results for clarity	5.25	5.88
Average of comprehension test	58.3	66.7

The results of the comparison between Expression 1, which follows the Japanese word order, and Expression 2, which uses the grammar of JSL and spatial/visual expressions, are shown in Figure 3 and Figure 4.

The average values of the survey ratings and comprehension test results are presented in Table I. Although only two participants indicated that they used JSL and the others were not familiar with JSL expressions, there were no significant differences in the rating values and comprehension test results for the two types of signs.

Additionally, in Expression 1, two participants scored zero on the comprehension test, whereas in Expression 2, no participant scored zero.

Even when technical terms and particles that connect words and phrases were explicitly indicated using fingerspelling and mouthing, the comprehension of the content presented in Japanese word order did not reach 60% (Expression 1).

In the follow-up survey regarding Expression 2, comments included the following:

- It is easy to form a visual image.
- There is a need to reconstruct the sentence in Japanese.

TABLE II. EXAMPLES OF NMMs OBSERVED DURING THE ANALYSIS OF NEWS SENTENCES.

Sentence type	NMMs
Topicalization	Eyebrows raised and eyes widened in the topic area at the beginning of the sentence
Wh-cleft sentence	Squinting and slightly shaking head in the middle of a sentence
Causal relationship	Eyebrow raised and head forward and fixed in the part of the condition
Complex sentence	Nodding motion before and after the clause

From these comments and observations during the verification, it was noted that while JSL expressions are understandable even for users who are not familiar with JSL and are more accustomed to Japanese, reconstructing Japanese sentences from sign language proves to be difficult, which impedes the writing process.

This confirms the need for a new way of writing sign language in professional situations.

IV. SIGN LANGUAGE FEATURES AND PROPOSED METHOD

To develop a new writing system, it is first necessary to identify the unique features of sign language sentences.

A. Analysis of News Texts

To address RQ2, an analysis was conducted to explore sign language-specific expressions in texts containing specialized content. Owing to the limited material on signed sentences expressing specialized content, sign language news was chosen for the analysis. News sentences typically employ topic-specific vocabulary and present factual information logically.

We analyzed 44 sentences from Sign Language News presented by four deaf news anchors at the Japan Broadcasting Corporation.

Table II shows examples of NMMs observed during the analysis. Topicalized sentences introduce the topic at the beginning, whereas wh-cleft sentences feature a question word in the middle.

B. Results of the Analysis of Signed Language Sentences

The analysis of signed sentences in Sign Language News yielded the following insights:

- Complex sentences were prevalent in sign language news texts (32 out of 44 sentences).
- Presenting the topic at the beginning of a sentence was frequently used (34 out of 44 sentences).
- Topicalization, wh-cleft sentence, and reason-for-sentence were used to introduce the topic.

Although Japanese sentences lacked topics, sign language sentences frequently present topics using sign language-specific expressions, such as topicalization/wh-cleft sentences or reason-for sentences, explicitly stating the reason at the onset of the sentence.

When using agreement verb When not using agreement verbs

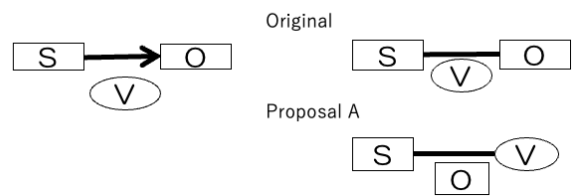


Figure 5. Improvements to the basic writing rule for sentences that do not use agreement verbs.

Thus, presenting a topic at the beginning of a sentence emerges as a sign language-specific feature crucial for facilitating comprehension by DHH individuals.

C. Proposed Method

To address RQ3, we proposed a new writing system that incorporates the identified sign-language-specific features.

Previously, we proposed a method for representing the spatial structure of sign language on a two-dimensional plane using symbols, such as the spatial representation of the subject and object [14]. After reviewing the basic rules, we consider a writing method that focuses on the macroscopic structure of sentences to highlight and visualize the topic in a manner conducive to DHH comprehension.

The rules of Proposal A for writing sign language are as follows:

Rule 1: The labels use the same text as the spoken language.

Rule 2: The subject and object are enclosed in squares, the predicate is enclosed in a circle, and the subject and predicate are connected by lines.

Rule 3: Clauses and phrases are represented using squares. In this manner, the hierarchical structure of a sentence can be visualized.

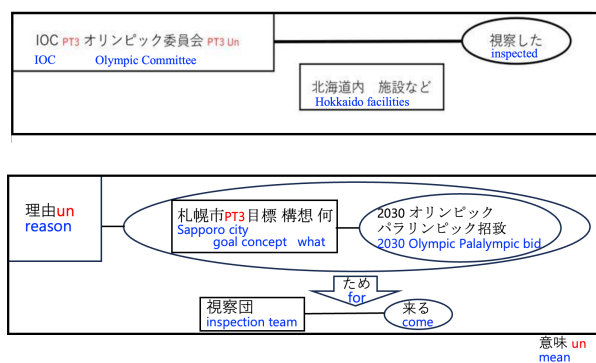
Rule 4: When reading a written sentence, it should be read from left to right and from top to bottom. SOV (Subject-Object-Verb) sentences follow the basic sentence structure for writing.

Rule 5: The upper-left square indicates topics, such as topicalization, wh-cleft sentence, reason-for-sentence.

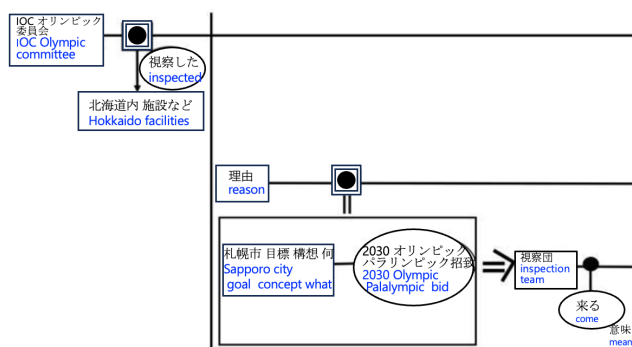
Rule 6: If there is a phrase that functions as a logical marker, such as “due to,” “reason for,” it should be written inside an arrow.

To understand specialized content, it is essential to understand the technical terminology used in textbooks and spoken languages accurately. Therefore, the same text as the spoken language is used as a label for writing sign language (Rule 1).

In our previously proposed rules, we reflected on the distinctive use of signing spaces in agreement verbs in the written form (Figure 5: original). However, because the basic word order in sign language is SOV, there was an opinion that it would be easier to write the following word order in cases where agreement verbs were not used. Therefore, we propose a writing method that follows the word order when agreement verbs are not involved (Figure 5: Proposal A).



(a) Proposal A



(b) Proposal B

Translation in English: The IOC Olympic member inspected facilities in Hokkaido. Because Sapporo City is aiming to bid for the 2030 Olympic and Paralympic Games.

Figure 6. Examples of a topicalization sentence and example of a reason sentence (Blue letters indicate translated English).

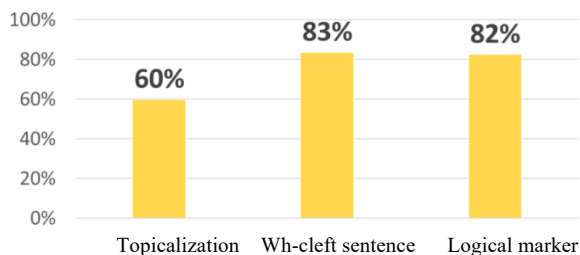
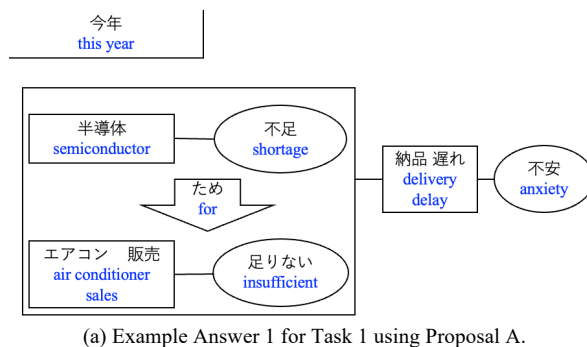
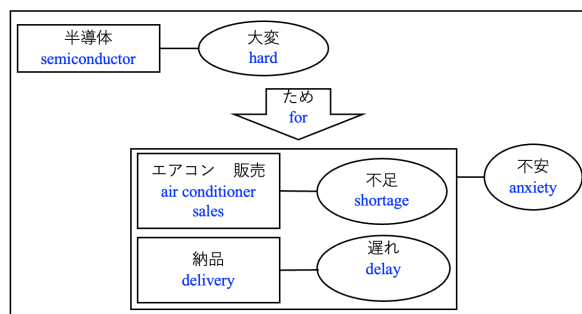


Figure 7. The accuracy rate of topicalization, wh-cleft sentences, and logical markers.



(a) Example Answer 1 for Task 1 using Proposal A.



(b) Example Answer 2 for Task 1 using Proposal A.

The English translation: This year, there were concerns about the shortage of air conditioners and delays in delivery due to the semiconductor shortage.

A list of sign labels: {this year} {semiconductor} {shortage} {due to} {air conditioners} {sales} {lack} {etc.} {delivery} {delay} {etc.} {concerns} {there were}

Figure 8. Examples of participants writing from a sign language video using the proposed method in Task1 (Reproduced from handwritten experimental results. Blue letters indicate translated English.).

Rules 5 and 6 notate structures specific to sign language sentences derived from the analysis of news scripts.

Examples of news scripts based on these rules are shown in Figure 6.

Figure 6(a) illustrates an example of a topicalized sentence using Proposal A. The sentence is enclosed in an outer-frame rectangle, with squares and circles representing the subjects, objects, and predicates. The rectangle in the top-left denotes the topic (Figure 6(a)).

In Proposal B, which employs a single line to preserve a word's position in the sign space across consecutive sentences, the branching point is surrounded by a double square to signify that the subject is the topic (Figure 6(b)).

V. EXPERIMENT

To verify the effectiveness of the proposed method, two experiments were conducted.

A. Experiment 1

1) Experimental Method for Experiment 1

We conducted an experiment to test the efficacy of the proposed sign language writing methods, specifically those based on Proposals A and B.

The participants were 12 university students who were either deaf or hard of hearing. Initially, the participants were briefed on the rules of the writing system and engaged in practice sessions to familiarize themselves with reading written signs using the proposed methods.

During the experiment, the participants were presented with a choice between Proposals A and B based on their preference for ease of understanding. They were then shown a video featuring a sign language news program.

Task 1: They were shown a video which contains only one sentence, without any ticker.

Task 2: They were shown a video in which the first and second sentences were accompanied by a ticker displaying only the main points. The third sentence was presented in sign language without tickers. The participants were instructed to transcribe the third sentence using their chosen writing method.

This setup aimed to simulate scenarios commonly encountered in academic settings, where signs are often displayed alongside textual materials, such as slides, allowing students to simultaneously view both sign language and written spoken language, such as English or Japanese.

Ethical approval was obtained from the Research Ethics Review of Tsukuba University of Technology, where the experiments were conducted.

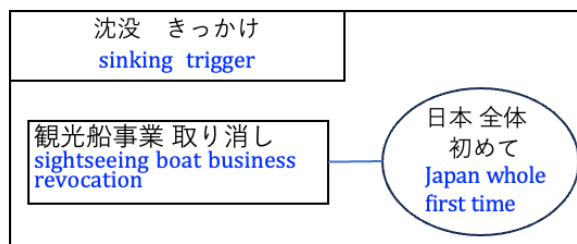
2) Experimental Results for Experiment 1

In the sign language news watching and writing experiment, 10 out of 12 participants opted for Proposal A, while two participants preferred Proposal B.

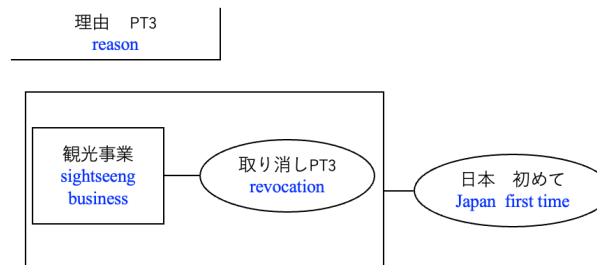
After a short practice session, participants answered whether each of the written signs contained a topic, logical marker, etc. The percentages of correct responses are shown in Figure 7. Logical markers and wh-splitting revealed a high percentage of correct answers, whereas the percentage of correct answers for topicalization was low. This is because it is difficult to distinguish between the case in which the subject of the sentence is surrounded by a square and the case in which it is the topic. Therefore, it was necessary to improve the topic by creating a double square.

In Task 1, when the ticker was not displayed, a problem arose where participants could not write because they did not know the labels for the sign language words. Therefore, in Task 1, when there were questions about the labels, the experimenter provided answers before transcription (e.g., semiconductor or air conditioner).

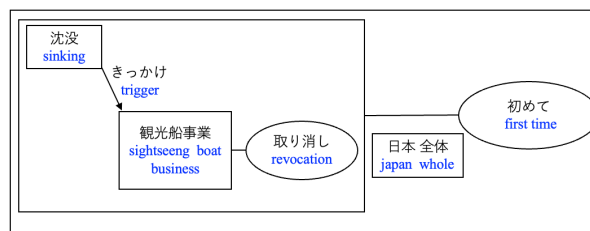
Figure 8 illustrates examples of the participants' writing in Task 1. The structure of a sentence is clearly visualized using logical markers. Nine out of ten people correctly used symbols for logical markers in the proposed method. However, different



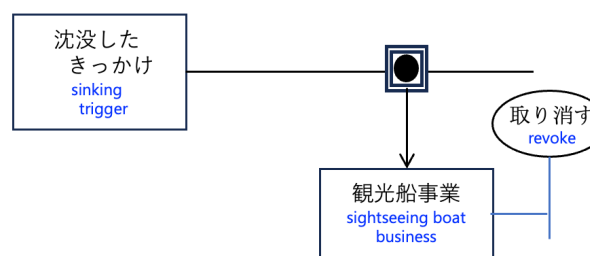
(a) Example Answer 1 for Task 2 using Proposal A. (An example where "revocation" was transcribed as a noun.)



(b) Example Answer 2 for Task 2 using Proposal A. (An example where "revocation" was transcribed as predicate.)



(c) Example Answer 3 for Task 2 using Proposal A. (An example of summarizing "a tourist boat business license was revoked due to sinking" into a single phrase.)



(d) Example Answer 1 for Task 2 using Proposal B.

Translation in English: This is the first time in the nation that a sightseeing boat business license has been revoked as a result of an accident.

A list of sign labels: {sinking} {trigger} {sightseeing} {boat} {business} {revocation} {Japan} {whole} {first time}

Figure 9. Examples of participants writing from a sign language video using the proposed method in Task 2 (Reproduced from handwritten experimental results. Blue letters indicate translated English.).

notations were observed in other parts of the sentence structure, indicating that the understanding of the sentences varied.

In Task 2, there were no questions regarding the labels for sign language words. Although there were multiple possible labels for a single sign word, 11 of the 12 participants opted for technical terms as their label. However, errors in symbol selection and placement were observed, presumably owing to the misinterpretation of sign language or the influence of the preceding context (Figure 9).

In Figure 9(a), “sinking trigger” is correctly selected as the topic, with the proposed symbol correctly employed. In Figure 9(b), there is a misreading of the sign word {reason}; however, the topic and pointing to the third person (PT3) were used as cues for structuring. Figure 9(c) shows that the topic is considered part of the phrase structure. Conversely, Figure 9(d) depicts the correct topic selection; however, errors in the placement of the symbols were observed. It can be inferred that the participants placed the topic in the subject position, possibly because of its placement at the beginning of the sentence.

B. Experiment 2

1) Experimental Method for Experiment 2

To clarify the challenges of writing sign language sentences, a deaf individual proficient in JSL read 44 sign language sentences and attempted to transcribe them by following predetermined rules. Assuming automatic transcription by computer, features such as topicalization, wh-question, pointing, and nodding were recorded and used as cues for transcription.

Following rules 1–6 for Proposal A, the procedure used for actual writing is as follows:

Step 1: When there is a topicalization marker, such as raised eyebrows, it is written as the topic in a rectangle touching the upper-left corner.

Step 2: When there is a wh-question, everything up to the wh word is written in a square touching the upper-left corner.

Step 3: For words indicating time and place, the label is written directly in the diagram as is.

Step 4: When there is a PT3, the word or the phrase up to that point is enclosed in a square as the subject or noun phrase.

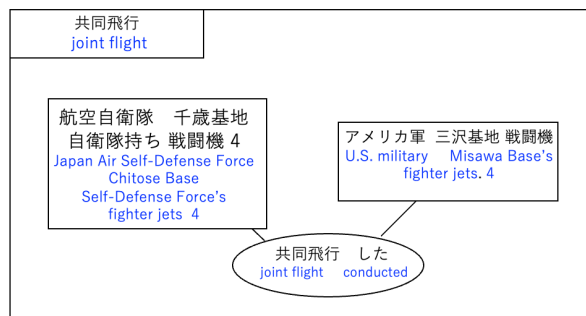
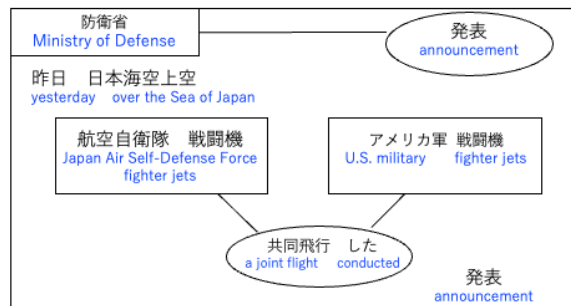
Step 5: If the phrase following the subject is an object, enclose it in a square; if it is a predicate, enclose it in a circle.

Step 6: Logical markers, such as “due to,” should be clearly indicated by writing them inside an arrow.

2) Experimental Results for Experiment 2

The experimental results identified the following challenges in structuring and writing sign language sentences using the proposed method:

- When the signing space is differentiated between the left and right, rules and procedures that reflect this are necessary (Figure 10(a)).
- There was a phrase considered a topic by PT3, even though there was no raised eyebrow marker (Figure 10(b)).
- In some cases, topicalization, wh-cleft, and logical markers were mixed in same sentence. Thus, the priority must be determined.



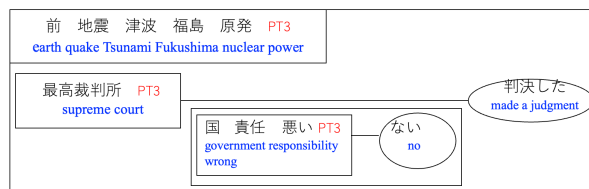
Translation in English:

(1st sentence) The Ministry of Defense announced that fighter jets from the Japan Air Self-Defense Force and the U.S. military conducted a joint flight over the Sea of Japan yesterday.
 (2nd sentence) It has been reported that the joint flight involved four F-15 fighter jets from the Japan Air Self-Defense Force's Chitose Base and four F-16 fighter jets from the U.S. military's Misawa Base.

A list of sign labels:

(1st sentence) {Ministry of Defense} {announcement} {yesterday} {the sea of Japan} {overhead} {Japan Air Self-Defense Force} {fighter jets} {U.S. military} {fighter jets} {joint flight} {finished} {announce}
 (2nd sentence) {a joint flight} {Japan Air Self-Defense Force} {Chitose Base} {Self-Defense Force's} {fighter jets} {4} {U.S. military} {Misawa Base's} {fighter jets} {4} {a joint flight} {conducted}

(a) An example illustrating the use of the signing space.



Translation in English: The Supreme Court has ruled that the government was not responsible for the Fukushima Daiichi nuclear power plant accident.

A list of signs labels:

{earth} {quake} {tsunami} {Fukushima} {nuclear} {power} {supreme court} {government} {responsibility} {wrong} {made} {judgement}

(b) An example where the phrase up to the PT3 at the beginning of the sentence is considered the topic.

Figure 10. Example that could not be written using only the predetermined steps.

- Phrases and clauses are also expressed through other clues, such as NMMs, and computers must accurately recognize such clues.

VI. DISCUSSION

In this section, the experimental results are analyzed in relation to the research questions.

A. RQ1: The need for a sign language writing system,

In the survey results of Verification 1, which compared explanations of mathematics using the Japanese word order for sign language words (Expression 1) and JSL grammar (Expression 2), all eight participants rated Expression 2 higher than Expression 1. A significant difference was observed in the average evaluation scores ($p < 0.01$).

Additionally, when comparing the lengths of the videos for Expressions 1 and 2, although Expression 2 included supplemental expressions not found in Expression 1, the videos for Expression 2 were shorter for all sign language sentences. In Expression 1, the use of finger spelling for technical terms made the videos longer, whereas in Expression 2, the information was conveyed more efficiently using spatial references and classifiers.

Through two verifications, in Expression 1, the participants were able to take notes in Japanese, but it was difficult in some cases for them to re-explain the content in their own words and score points on the comprehension test.

However, in Expression 2, the meaning was conveyed as an image, but it was difficult to write because the participants did not know the technical terms needed for writing, making it challenging to obtain high scores on comprehension tests.

In other words, it was confirmed that although it is easy to understand explanations using the grammar of JSL, it is difficult to write them in written Japanese, and a method of writing sign language is needed.

B. RQ2: Sign Language-Specific Features

When comparing the two videos created for Verification 1 and Verification 2, the following differences were observed:

- [Explicit subject] In Japanese (Expression 1), the subject is omitted, whereas in the JSL (Expression 2), it is sometimes explicitly stated.
- [Additional explanations] In Expression 2, there were supplemental expressions and repetitions that were not present in Expression 1.
- [Use of CL] In Expression 1, technical terms were expressed using fingerspelling and mouthing, whereas in Expression 2, CL (classifiers), which are sign language elements that express the characteristics of objects and movements with hand shapes, were used extensively.

To clarify the structural characteristics of more sign language sentences, news scripts were analyzed.

Comparison between Japanese and signed news sentences revealed the following features:

- Complex sentences were often used, with over 70% of the sentences exhibiting complexity, contrary to the common belief that sign sentences are short and simple.
- The structuring of complex sentences in sign language often involves presenting the topic at the sentence outset.

Sign language employs specific expressions such as topicalization and the wh-cleft to introduce and emphasize topics. For instance, in sentences indicating reasons, sign language presents the word “reason” at the beginning, followed by the logical marker “for,” and conclude with a phrase expressing the result, a structure not mirrored in Japanese (Figure 6(a)).

These specific expressions are considered to aid in conveying technical concepts in a digestible manner for DHH individuals.

C. RQ3: Writing Sign Language Sentences

Developing a writing system for sign language requires consideration of the perceptual characteristics of DHH individuals and their information processing. Therefore, such a system must incorporate spatial representations, time-series depictions, and the visualization of grammatical and logical structures.

We propose a method that projects spatial and time series representations onto a 2D (two-dimensional) plane, and uses symbols to represent grammatical and logical structures. In addition to basic spatiotemporal representation, our approach focuses on the macroscopic structure of sentences, represented by NMMs and other visual cues.

The experimental preference for Proposal A by 10 of the 12 participants underscores that emphasizing the topic at the beginning is effective. Topic sentences are represented by NMMs such as raised eyebrows (Table II). Although NMMs are said to be challenging for learners to master, written signed sentences can aid in comprehending these expressions.

Regarding sign labels, 11 of the 12 participants used technical terms in their real-time sign writing. To use technical terms as labels, we must consider how sign language and slides are presented.

Although the explanation and practice of the proposed method were brief, in Task 1, six out of ten participants were able to write a topicalization and nine out of ten participants were able to write a logical marker. This was the same trend as the comprehension of the reading of the structure of written signs (Figure 7). It is necessary to highlight topicalization with a double rectangle.

In this experiment, the participants did not necessarily consider the structure of the whole sentence before writing it, but rather tended to record the sign labels in the order of the time series.

The experiment revealed difficulties in selecting and positioning symbols (Figure 9(d)), highlighting the need for computer support such as automatic placement and insertion of symbols.

D. RQ4: Challenges of the proposed sign language writing method

- 1) Sign language labels as technical terms

In Experiment 1, it was found that by displaying sign language and tickers simultaneously, users could write down technical terms. In specialized explanatory contexts, it is assumed that sign language will be used alongside slides and it will be necessary to identify labels from a larger amount of text.

For instance, when fingerspelling is used, it can be used to identify technical terms [15], or when pointing gestures are made, the corresponding technical terms can be extracted. In such cases, computer support is considered to be effective.

2) To identify sentence structure

In the actual transcription experiment, there were many variations in the choice of symbols and positioning of each label. Many of the experiment participants were not familiar with JSL, and it is thought that many NMMs were overlooked.

However, we found that even users unfamiliar with JSL were able to write, read, and distinguish the macrostructures of sentences, such as topicalization and logical markers, using the proposed method after a short practice.

Although the basic word order in JSL is typically SOV, word order can be changed to present a topic, and the first position in a sentence does not always indicate the subject. Moreover, in JSL, there are homonyms between nouns and verbs, which make it difficult to distinguish between similar signs.

For example, In Figure 9(a)(b)(c)(d), there are some notational distortions between the verb “revoke” and the noun “revocation,” and between the subject enclosed in a square and the predicate enclosed in a circle.

For users unfamiliar with JSL, we tried a writing procedure using relatively easy-to-recognize NMMs, such as pointing (PT3), nodding (Un), and raising eyebrows. In the procedure we envisioned, PT3 was used to identify the subject noun or noun phrase. However, because PT3 is frequently used for other meanings, ambiguity remained.

Additionally, clauses and phrases can be represented by other NMMs, such as head movement, head position, and the timing of the nod [16]. Therefore, it is necessary to add rules that use NMMs to identify sentence structures in specialized texts.

To create a computer-assisted sign language writing system, it is important not only to properly recognize technical terms in the signed text but also to properly discriminate these NMMs. It is necessary to sequentially incorporate known areas, such as the differences between nods that mark the boundaries of a clause and other nods.

The sentence structures in Figure 8(a) and (b) are different because the grammar of JSL cannot be read. In other words, transcribing sign language sentences can be viewed as a way to visualize the level of understanding of sign language content.

3) Reflects the signing space

When the signing space is used effectively, it is considered appropriate to reflect it in the writing (Figure 10(a)).

When compared to the Japanese transcription in Figure 11, Figure 10 clearly distinguishes the roles of the left and right



Translation in English:

(1st sentence) The Ministry of Defense announced that fighter jets from the Japan Air Self-Defense Force and the U.S. military conducted a joint flight over the Sea of Japan yesterday.

(2nd sentence) It has been reported that the joint flight involved four F-15 fighter jets from the Japan Air Self-Defense Force's Chitose Base and four F-16 fighter jets from the U.S. military's Misawa Base.

Figure 11. In the case of writing a sentence with sign language words arranged in Japanese word order.

and maintains these roles consistently between the first and second sentences, making the sentences easier to understand.

To automatically perform such a transcription, it is necessary to prioritize the identification and representation of the signing space when arranging the layout.

E. Limitation

The limitations of this study include the small number of participants, variability in sign language proficiency, and the limited number of signed sentences. Further research with a larger number of expressions and sentence patterns is required to design a system that is useful for improving the learning performance of DHH individuals.

VII. CONCLUSION AND FUTURE WORK

This study aimed to develop a computer-assisted writing system tailored to the perceptual characteristics of DHH individuals by considering the visual and spatial nature of sign language and the unique characteristics of signed sentence.

First, to examine the necessity of such a writing system, we conducted verifications while conveying specialized content using two types of sign language expressions. The

verifications revealed that while JSL expressions effectively convey the image and are highly rated by DHH users, the inability to write them down presents challenges, indicating the need for a system to write sign language.

By analyzing news sentences in sign language, we confirmed that numerous expressions specific to sign language, such as topicalized and wh-cleft sentences, were used. To establish a new method of expression that is intuitive and understandable for the deaf, we proposed a writing system that reflects these features and conducted an experiment in which participants wrote a sign language video.

The results of the experiment demonstrated that by using the proposed method, participants could write signed sentences with sign language-specific features.

Furthermore, to represent the structure of sign language sentences of specialized contents, it is necessary to recognize not only the presence or absence of raised eyebrows, nodding, and pointing is important, but also spatial expressions, the intensity of pointing and nodding, and elements such as eye movement and head orientation.

In addition, writing signs while maintaining the structure of the signed sentence leads to the visualization of each student's comprehension of the content and is expected to be applied to learning in the future, such as checking comprehension independent of the power of the spoken language.

In future, we intend to expand our research by conducting a broader survey involving a larger sample of sentences. This will enable us to further refine our proposed writing system and provide support for communication and learning among DHH individuals.

ACKNOWLEDGMENT

We express our sincere gratitude to Dr. Takaaki Arai for their invaluable advice throughout the course of this study.

This study was supported by JSPS KAKENHI (grant numbers JP22K02999, JP19K11411 and JP24K14243). We would like to thank Editage (www.editage.jp) for English language editing.

REFERENCES

- [1] N. Kato, Y. Nameta, A. Shitara, and Y. Shiraishi, "Sign Language Writing System: Focus on the Representation of Sign Language-Specific Features," The 17th International Conference on Advances in Computer-Human Interactions (ACHI2024) IARIA, May. 2024, pp. 188-192, ISBN: 978-1-68558-163-3.
- [2] L. Berke, S. Kafle, and M. Huenerfauth, "Methods for evaluation of imperfect captioning tools by deaf or hard-of-hearing users at different reading literacy levels," Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, ACM, Apr. 2018, pp. 1-12. doi: 10.1145/3173574.3173665.
- [3] K. R. Miller, "American Sign Language: Acceptance at the university level," *Language, Culture and Curriculum*, vol. 21, no. 3, pp. 226-234, Nov. 2008, doi: 10.1080/07908310802385899.
- [4] WFD, "7 September 2016: WFD Position Paper on the Language Rights of Deaf Children," [Online]. Available from: <http://wfdeaf.org/news/resources/wfd-position-paper-on-the-language-rights-of-deaf-children-7-september-2016/> 2024.11.30
- [5] M. Huenerfauth and V. Hanson, "Sign language in the interface: access for deaf signers," in *The Universal Access Handbook*, CRC Press, pp. 1-18, 2009. doi: 10.1201/9781420064995-c38.
- [6] J. Napier, "University Interpreting: Linguistic Issues for Consideration," *Journal of Deaf Studies and Deaf Education*, vol. 7, no. 4, pp. 281-301, Oct. 2002, doi: 10.1093/deafed/7.4.281.
- [7] H. G. Lang et al., "A study of technical signs in science: implications for lexical database development," *Journal of Deaf Studies and Deaf Education*, vol. 12, no. 1, pp. 65-79, Aug. 2006, doi: 10.1093/deafed/enl018.
- [8] D. A. Grushkin, "Writing Signed Languages: What For? What Form?," *American Annals of the Deaf*, vol. 161, no. 5, pp. 509-527, 2017, doi: 10.1353/aad.2017.0001.
- [9] T. Hanke, "HamNoSys – Representing sign language data in language resources and language processing contexts," *Proceedings of the 1st Workshop on the Representation and Processing of Sign Language*, 2004, pp. 1-6.
- [10] H. Van Der Hulst and R. Channon, "Notation systems," in *Sign Languages*, 1st ed., D. Brentari, Ed., Cambridge University Press, pp. 151-172, 2010, doi: 10.1017/CBO9780511712203.009.
- [11] C. Valli and C. Lucas, *Linguistics of American Sign Language: an introduction*, 3rd ed. Washington, D.C: Gallaudet University Press, 2000.
- [12] E. Rathkey, "Can ASL-gloss be used as an instructional tool to teach written English to the deaf?," *Open Access Dissertations*, paper 835, Jan. 2019, doi: 10.23860/diss-rathkey-emma-2019.
- [13] K. Yano and A. Utsumi, "Pipeline signed Japanese translation using PBSMT and transformer in a low-resource setting," *Journal of Natural Language Processing*, vol. 30, no. 1, pp. 30-62, 2023, doi: 10.5715/jnlp.30.30.
- [14] N. Kato, Y. Hotta, A. Shitara, and Y. Shiraishi, "Visually-structured written notation based on sign language for the deaf and hard-of-hearing," *Proceedings of the 15th International Conference on Computer Supported Education - Volume 2: CSEdu, INSTICC, SciTePress*, 2023, pp. 543-549. doi: 10.5220/0011988700003470.
- [15] S. Tanaka, A. Okazaki, N. Kato, H. Hino, K. Fukui, "Spotting fingerspelled words from sign language video by temporally regularized canonical component analysis". *IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, 2016, pp. 1-7. doi: 10.1109/ISBA.2016.7477238.
- [16] N. Shimotani, "Analyzing head nod expressions by L2 learners of Japanese Sign Language: A comparison with native Japanese Sign Language signers," in *East Asian Sign Linguistics*, Kazumi Matsuoka, Onno Crasborn and Marie Coppola, Ed., Berlin, Boston: De Gruyter Mouton, 2023, pp. 241-262. doi: 10.1515/9781501510243-009.

Regression Model-Based Prediction for Building Energy Star Score of New York City

Fan Zhang, Baiyun Chen*, Fan Wu,
Faria Brishti, Sameeruddin Mohammed
Computer Science Department
Tuskegee University
Tuskegee, USA

e-mail: {fzhang9458;bchen;fwu;
fbrishti7995;smohammed8703}@tuskegee.edu

Ling Bai
School of Engineering
The University of British Columbia
Kelowna, Canada
e-mail: ling.bai@ubc.ca

Abstract—Machine learning algorithms have recently shown promise in predicting Energy Star Scores for buildings, outperforming traditional forecasting methods. While previous studies have focused on specific building types, this comprehensive research expands the scope to analyze and predict Energy Star Scores across four diverse building categories in New York City: residential, educational, commercial, and lodging structures. Our study employs rigorous feature engineering and selection to develop nine distinct regression models applied to these four building types. We compare various machine learning algorithms to identify the most effective predictive model for each category. The Gradient Boosting Regressor (GBR) consistently emerges as the top performer across building types, demonstrating superior accuracy and stability in predictions. We provide a detailed analysis of feature importance for each building category, offering insights into the key factors influencing energy efficiency across different sectors. By extending the analysis to multiple building types and employing a range of regression models, this study contributes to a more comprehensive understanding of urban energy efficiency and provides tailored strategies for improving energy performance across New York City's diverse building stock for urban planners, building managers, and policymakers.

Keywords—Machine learning; Regression; Data analysis; Model evaluation.

I. INTRODUCTION

This paper serves as an expansion of our initial study on the prediction of the Energy Star Score for residential buildings: A case study of New York City [1]. In this expanded version, we broaden the scope of our regression models to forecast Energy Star Scores across a diverse range of building types, encompassing not just residential structures but also educational buildings, commercial buildings, and lodging buildings in New York. All the buildings studied in this paper possess a common attribute, namely, that individuals spend substantial amounts of time within them.

As economic and social development has progressed, the consumption of energy and water resources by human behaviors has increased by an order of magnitude, leading to a rise in annual carbon dioxide emissions and a severe reduction of water resources [2]. This trend has significant implications for the sustainable development of human society. Buildings account for approximately 40% of global energy consumption,

a percentage projected to increase in the coming decades [3]. This growth is attributed to two main factors: the frequent extreme temperature fluctuations caused by climate change [4], and rising human demands for housing and improved living standards [5]. Among various energy end uses in buildings, space heating typically consumes the largest share, accounting for over 30% of total energy use, which is followed by water heating, cooling, ventilation, and lighting, though the exact order can vary depending on the building type [6]. Notably, residential buildings are responsible for almost 70% of the energy consumption of the sector, mainly due to the usage for cooking and heating [7]. Fortunately, it illustrates a great potential to enhance the energy efficiency of buildings by analyzing the retrofit options or adjusting human activities in energy consumption.

The Energy Star Score for buildings was developed by the United States Environmental Protection Agency (EPA) in collaboration with the U.S. Department of Energy (DOE), evolving from the broader Energy Star Program launched in 1992 [8]. This 1-100 scoring system provides a standardized method for measuring and comparing energy efficiency across different types of buildings. A score of 100 indicates top performance, placing the building among the most energy-efficient nationwide, while a score of 1 represents the lowest performance [9]. The Energy Star Score is a crucial metric for assessing the energy efficiency of buildings, enabling stakeholders to evaluate and compare building performance objectively. Estimating this score is therefore essential for building owners, managers, and policymakers seeking to improve energy efficiency in the built environment.

However, the complexity of building energy consumption, influenced by numerous factors such as weather conditions, occupancy patterns, building characteristics, and operational schedules, etc. poses significant challenges to accurately predicting building energy consumption, which directly influences the Energy Star Score. A lot of efforts from academia, industry, and governments have originated multiple methods or tools for the estimation of buildings' energy consumption. The Building Energy Software Tools Directory [10] provides comprehensive information on building software tools for evaluating energy efficiency and sustainability in buildings. This directory also

*Corresponding author.

shows that efforts can be derived for different components to minimize energy consumption. With the widespread application of machine learning techniques, a growing number of researchers have recently proposed to introduce regression models for predicting building energy consumption, offering a data-driven method to navigate this intricate landscape of variables and their interactions [11]–[13]. Linear regression model is the most basic model applied to predict building energy consumption, due to its simplicity, straightforward implementation, and computational efficiency [14]. However, linear regression often falls short in capturing the intricate, non-linear relationships between input variables and energy consumption outcomes. Thus, regression models capable of handling non-linear relationships are often necessary to achieve higher prediction accuracy in the multifaceted domain of building energy consumption.

Various advanced regression techniques have been proposed to address the limitations of linear regression in predicting building energy consumption. Jung et al. and Ma et al. suggested using support vector regression (SVR) due to its ability to handle complex non-linear relationships in data [15], [16]. Yu et al. proposed tree-like structures, particularly decision trees, to analyze building parameters and predict energy demand, allowing for the identification of key influencing factors [17]. To enhance the performance of single decision trees, the Random Forest method was introduced, which ensembles multiple trees [18]. This concept of ensembling improves predictive performance by combining multiple models together to leverage their collective strengths, reduce individual weaknesses, and capture diverse aspects of the data. Similar principles are employed in Gradient Boosting and extreme gradient boosting models, both of which have been applied to building energy consumption prediction [19], [20].

Artificial Neural Networks (ANN), inspired by biological neural networks, have gained popularity for their ability to solve non-linear problems associated with high-dimensional datasets [21]. Deep Learning, an advanced form of ANN, excels at capturing consumption patterns from historical data and discovering non-linear relationships between inputs and outputs [22]. Among the various types of neural networks, the Multilayer Perceptron (MLP) has emerged as a particularly effective tool for predicting building energy consumption, including heating and cooling loads [23]. This application represents a rapidly growing research area due to its potential to significantly enhance energy efficiency in building management systems. However, these methods, including Support Vector Machines, Decision Trees, Random Forest, and Artificial Neural Networks, often require significant computational resources for parameter optimization and model tuning. Deep Learning, in particular, demands not only substantial computing power but also high-quality, large-scale labeled datasets.

In contrast, some researchers have explored simpler methods like k -Nearest Neighbors (k NN) for building energy consumption prediction. k NN forecasts energy consumption by identifying similar past instances based on relevant factors such as weather conditions, appliance usage, and time of day [24].

This method's appeal lies in its simplicity, ease of interpretation, and minimal assumptions about data distribution, with only one parameter (k) to optimize.

This study focuses on urban buildings, given the high population density and concentrated energy consumption in metropolitan areas. Almost all the aforementioned regression models are employed to forecast the Energy Star Score of buildings using disclosed energy and water consumption data from New York City. The performance of these various approaches is then systematically compared. Moreover, by evaluating the significance of different features, this research identifies key factors that substantially influence energy consumption for each building type. These insights offer valuable guidance for future building design, retrofit strategies, and occupant behaviors related to heating and cooking. Ultimately, this work aims to support efforts to reduce emissions and conserve energy in urban environments, contributing to more sustainable and efficient city infrastructures.

The structure of the paper is as follows. Section II briefly introduces the five conventional regression methods utilized in this work. Section III depicts the modeling procedure and results for the residential building energy consumption data in New York, presenting and discussing the findings. We conclude with Section IV.

II. METHODS

Regression approaches, one of the most popular types of machine learning algorithms, demonstrate superior predictability with promising results in various domains, including energy consumption [25], bankruptcy prediction [26], air pollution [27], epidemiology [28], and some other applications. This study introduces 9 typical regression methods, including k -Nearest Neighbor Regression [29], Linear Regression [30], Ridge Regression [31], Decision Tree Regression [32], Random Forest Regression [33], Support Vector Regression [34], and Gradient Boosting Regression [35], eXtreme Gradient Boosting Regression [36], and Multi-Layer Perceptron [37] to predict the Energy Star Score of residential buildings and investigates the prediction results using four metrics, i.e., MAE, SSE, R^2 , Adjusted R^2 [38]. The coefficient of determination, R^2 , measures the proportion of the variance in the dependent variable that is predictable from the independent variables. Adjusted R^2 is a modified version of R^2 that adjusts for the number of predictors in the model.

Mathematically, given a training dataset D with features X and target values Y , and a new data point x for which we want to predict the target value \hat{y} , we briefly introduce the nine regression models and calculate \hat{y} in each regression model accordingly.

A. k -Nearest Neighbor Regression

k NN regression, or k -Nearest Neighbors regression, is a non-parametric regression method that predicts target values by averaging the observed values of the k nearest samples in the feature space [29]. Hence,

$$\hat{y} = \frac{1}{k} \sum_{i=1}^k y_i, \quad (1)$$

where y_i are the target values of the k nearest neighbors of \mathbf{x} . The nearest neighbors are typically determined based on a distance metric, such as Euclidean distance.

B. Linear Regression

Linear regression is a parametric regression technique that models the linear relationship between dependent and independent variables by minimizing the residual sum of squares [30]. The predict value \hat{y} is calculated using (2):

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n, \quad (2)$$

where $\beta_0, \beta_1, \beta_2, \dots, \beta_n$ are the estimated parameters for the linear regression model and x_1, x_2, \dots, x_n are the values of the independent variables for the new data point.

C. Ridge Regression

Ridge Regression is a regularized linear regression method that introduces an L2 penalty term to mitigate multicollinearity and reduce model variance [31]. The prediction \hat{y} is calculated using:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n, \quad (3)$$

where $\beta_0, \beta_1, \beta_2, \dots, \beta_n$ are the estimated parameters. These parameters are obtained by minimizing:

$$\sum_{i=1}^m (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^n \beta_j^2, \quad (4)$$

where λ is the regularization parameter that controls the strength of the penalty.

D. Decision Tree Regression

Decision Tree Regression is a non-parametric model that predicts target values by recursively partitioning the feature space into regions with homogeneity and assigning predictions based on local sample averages [32]. The prediction \hat{y} for a new data point \mathbf{x} is given by:

$$\hat{y} = \sum_{m=1}^M c_m I(\mathbf{x} \in R_m), \quad (5)$$

where M is the number of leaf nodes, c_m is the predicted value in the m -th leaf node, R_m is the region of feature space corresponding to the m -th leaf node, and I is an indicator function that equals 1 if \mathbf{x} is in region R_m and 0 otherwise.

E. Random Forest Regression

Random Forest Regression is an ensemble learning approach that combines predictions from multiple decision trees trained on bootstrapped samples to reduce variance and improve generalization [33]. \hat{y} is predicted by (6):

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N f_i(\mathbf{x}), \quad (6)$$

where $f_i(\mathbf{x})$ is the prediction of the i^{th} decision tree for the new data point \mathbf{x} and N is the total number of decision trees in the Random Forest.

F. Support Vector Regression

Support Vector Regression is a regression technique that seeks a hyperplane in a high-dimensional space to minimize prediction errors within a predefined tolerance, supported by a margin [34]. \hat{y} is predicted by (7):

$$\hat{y} = \mathbf{w}^T \cdot \mathbf{x} + b, \quad (7)$$

where \mathbf{w} is the weight vector and b is the bias term.

G. Gradient Boosting Regression

Gradient Boosting Regression is a sequential ensemble method that optimizes a differentiable loss function by constructing regression trees in a stage-wise manner using gradient descent in the function space [35]. \hat{y} is predicted by (8):

$$\hat{y} = \sum_{i=1}^N \gamma_i f_i(\mathbf{x}) \quad (8)$$

where γ_i is the learning rate that controls the contribution for each learner, $f_i(\mathbf{x})$ is the prediction of the i^{th} decision tree for the new data point \mathbf{x} and N is the total number of decision trees in the Gradient Boosting model.

H. eXtreme Gradient Boosting Regression

XGBoost is a highly efficient gradient boosting implementation that integrates advanced regularization techniques and parallel processing to enhance computational performance [36]. The prediction \hat{y} is given by:

$$\hat{y} = \sum_{k=1}^k f_k(\mathbf{x}), \quad (9)$$

where k is the number of trees, f_k represents the k -th tree. The objective function to be minimized is:

$$\mathcal{L} = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k), \quad (10)$$

where l is a differentiable convex loss function and Ω is the regularization term.

I. Multi-Layer Perceptron

Multi-Layer Perceptron (MLP) is a fully connected feed-forward neural network that approximates complex functions through layered processing and non-linear transformations [37]. For a MLP with L layers, the prediction \hat{y} is calculated as:

$$\hat{y} = f_L(W_L \cdot f_{L-1}(W_{L-1} \cdot \dots f_1(W_1 \cdot \mathbf{x} + b_1) \dots + b_{L-1}) + b_L), \quad (11)$$

where W_l and b_l are the weight matrix and bias vector for layer l respectively, and f_l is the activation function for layer l . Common choices for f_l include ReLU, sigmoid, and tanh functions.

J. Performance Metrics

Four commonly used performance metrics are employed in this work. They are Mean Absolute Error (MAE), Sum of Squared Errors (SSE), Coefficient of Determination (R-squared, R^2), and Adjusted R^2 . MAE measures the average absolute difference between the predicted values and the actual values; SSE measures the total squared difference between the predicted values and the actual values; R^2 can be interpreted as the percentage of the variance in the dependent variable that is explained by the independent variables; Adjusted R^2 provides a more accurate assessment, which penalizes the addition of unnecessary variables to the regression model [39].

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (12)$$

$$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (13)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (14)$$

$$\text{Adjusted } R^2 = 1 - \frac{(1 - R^2) \cdot (n - 1)}{n - k - 1} \quad (15)$$

These performance measures aid in evaluating the quality of fit and accuracy of regression models, facilitating the comparison and assessment of various models and their capacity for prediction.

III. CASE STUDY

Predicting the Energy Star Score follows the standard machine learning workflow, which consists of four stages: data collection, data preprocessing, model training, and model testing, as shown in Figure 1 [11]. Data collection gathers crucial building and energy consumption data. Data preprocessing involves cleaning and preparing data for analysis. Model training consists of selecting algorithms, setting parameters, and training the models. Finally, in the model testing stage, it examines the models' ability to predict the Energy Star Score.

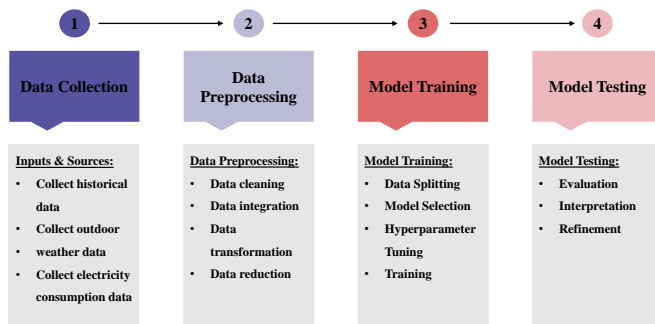


Figure 1. Workflow of predicting building Energy Star Score.

Data used for the regression prediction corresponds to the energy and water data disclosed for Local Law 84 of the New York City in the calendar year 2021 [40]. It encompasses a

diverse range of building types, including residential buildings, educational buildings, commercial buildings, lodging buildings, factories, cultural institutions, and various other structures. In this study, we concentrated on four specific building types: multifamily residential buildings, K-12 schools, office buildings, and hotels. We chose these categories because they have significantly more available data compared to other types of buildings. A richer data set is advantageous in constructing a more robust predictive model and mitigating the uncertainty caused by limited data. After cleaning the dataset by removing rows with missing values and outliers, we extracted a total of 13,871 records from the original 22,479 rows, focusing on our four selected building types. This cleaned dataset comprises 10,802 records for multifamily residential buildings, 1,564 records for K-12 schools, 1,112 records for office buildings, and 393 records for hotel buildings. This substantial sample, representing about 62% of the original data, provides a robust foundation for our predictive models across these key urban building categories.

The original dataset comprises 249 columns, with the Energy Star Score column serving as the target variable for prediction. The score quantifies the property's performance relative to similar ones, rated on a scale of 1 to 100, where 1 denotes the poorest-performing buildings, and 100 indicates the best-performing ones. The remaining columns are considered as variables constituting the potential features in the regression model. A comprehensive explanation for each column can be found in the data dictionary [40].

Given that these four building types belong to distinct categories with varying energy consumption patterns, occupancy behaviors, and building functions, developing a single model to predict Energy Star Scores across all categories could lead to underfitting. The significant differences in sample sizes among the categories further complicate this issue. To address these challenges and to better capture the unique energy consumption characteristics of each building type, we opted to develop separate regression models for each category, which helps to maximize prediction accuracy by tailoring each model to the specific features and patterns of its respective building type.

A. Feature Statistics

Prior to constructing the predictive model for residential energy consumption, it is imperative to thoroughly explore the features within the original dataset. As it is known, each feature holds varying degrees of importance, with the Energy Star Score column being the most crucial as it serves as the target variable for prediction. Therefore, we first use a histogram to represent the distributions of this target variable, as shown in Figure 2.

Figure 2 illustrates the distribution of Energy Star Scores across four different building types: multifamily housing, K-12 schools, offices, and hotels. Each subfigure corresponds to one type separately. Notably, none of these distributions conform to either a uniform or a normal distribution. Multifamily housing in Figure 2(a) shows high frequencies at both ends with lower, uneven distribution in the middle. Both K-12 schools in Figure 2(b) and office buildings in Figure 2(c) exhibit

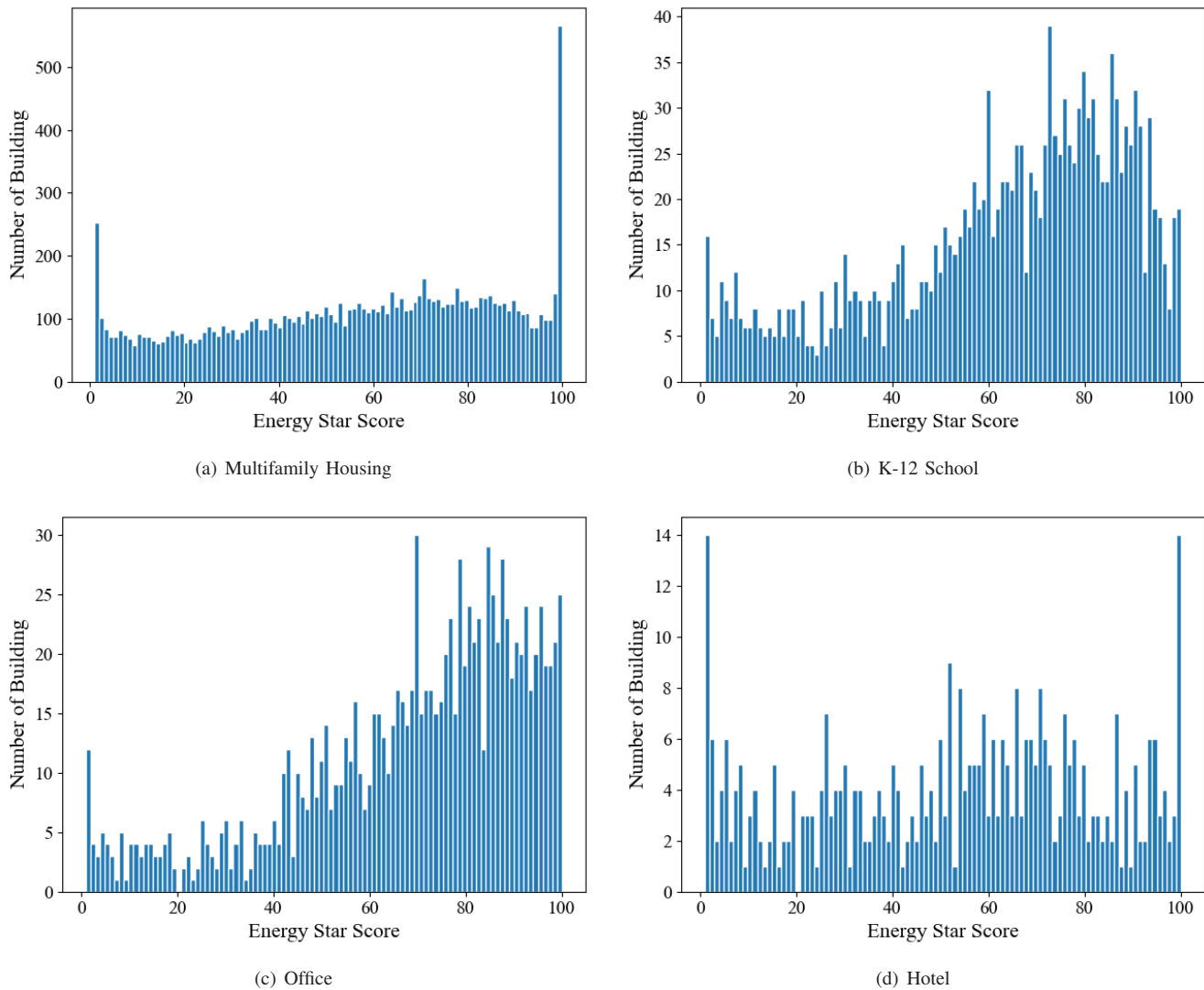


Figure 2. Distribution of Energy Star Score across four types of buildings.

similar patterns that can be described as slightly right-skewed bimodal distributions. They show a small peak in the lower score range (around 0-20), with scores gradually increasing towards the higher end, reaching a maximum peak near 100. This suggests that while a significant portion of these buildings achieve high energy efficiency, there's also a smaller group with lower efficiency. Hotels in Figure 2(d) present the most irregular pattern, with scores scattered across the range and multiple local peaks. These diverse, non-standard distributions across all building types underscore the complexity of energy performance in different sectors and highlight the necessity for advanced regression techniques rather than traditional statistical methods for accurate modeling and prediction of Energy Star Scores.

Next, we need to screen out the more important variables to the target variable for modeling from the 248 features, a step commonly known as feature selection. This process stands as one of the pivotal stages in the entire machine learning

workflow. The efficacy of a machine learning model heavily relies on the predictive capability of the selected features. Even a simple linear model can showcase commendable performance if these features exhibit strong predictability. Conversely, the modeling process should exclude features with weaker predictive power. Their inclusion would not only increase model complexity but also compromise prediction accuracy.

In this study, we employ a non-parametric statistical technique, Kernel Density Estimation (KDE), to assess the effect of various variables on the distribution of the target variable. Variables demonstrating substantial fluctuations in the distribution of energy scores across different values are deemed significant, whereas those exhibiting minimal variation are deemed inconsequential. For example, we explore the impact of districts on the distribution of the Energy Star Score, as illustrated in Figure 3. We first categorize the datasets into different groups based on five districts in New York: Bronx,

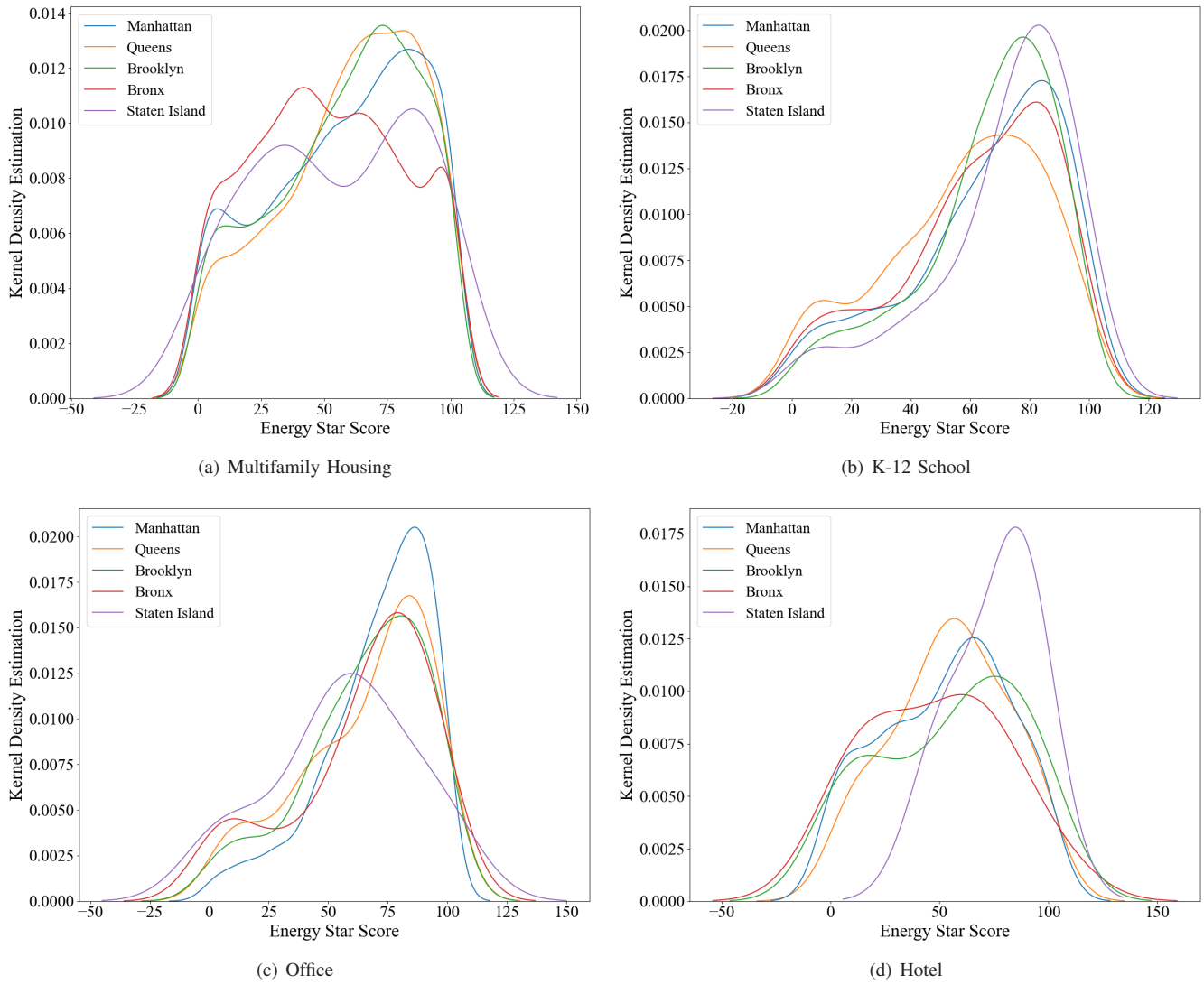


Figure 3. Distribution of Energy Star Score in different districts.

Manhattan, Brooklyn, Queens, and Staten Island, then we employ the Gaussian Kernel function to smooth the probability density estimation of different groups.

The KDE analysis across the four types reveals that the distribution of Energy Star Scores is generally consistent across the five districts, as shown in Figure 3. For all building types, the score distributions show similar patterns, with peaks around the same ranges. The analysis reveals that within each building category (multifamily, office, educational, and lodging), the Energy Star Score distributions show similar patterns across all five districts of New York City. This suggests that a building’s geographical location within the city does not significantly influence its energy performance when compared to other buildings of the same type. Therefore, the district variable was not included as a predictor in our final models. Although Staten Island displays a somewhat distinct pattern for hotel buildings in Figure 3(d), this deviation is attributed to the fact that there are only three samples from this district,

which is statistically insufficient to accurately represent the true distribution. Consequently, the district variable is not recommended for inclusion in the modeling process due to its limited contribution to predictive accuracy. This insight can help streamline future models and focus attention on variables that demonstrate greater discriminative power in the context of building energy efficiency.

Subsequently, we conduct correlation analysis to detect multicollinearity in two or more independent variables that are highly correlated with each other, possibly resulting in instability and inflated standard errors in regression models. By identifying and removing highly correlated variables, we can mitigate multicollinearity and improve the stability and interpretability of the model.

Figure 4 demonstrates the correlation analysis result of “Site EUI” and “Weather Normalized Site EUI” in the scatter diagram for four building types. EUI refers to the Energy Use Intensity, which measures the ratio of actual energy

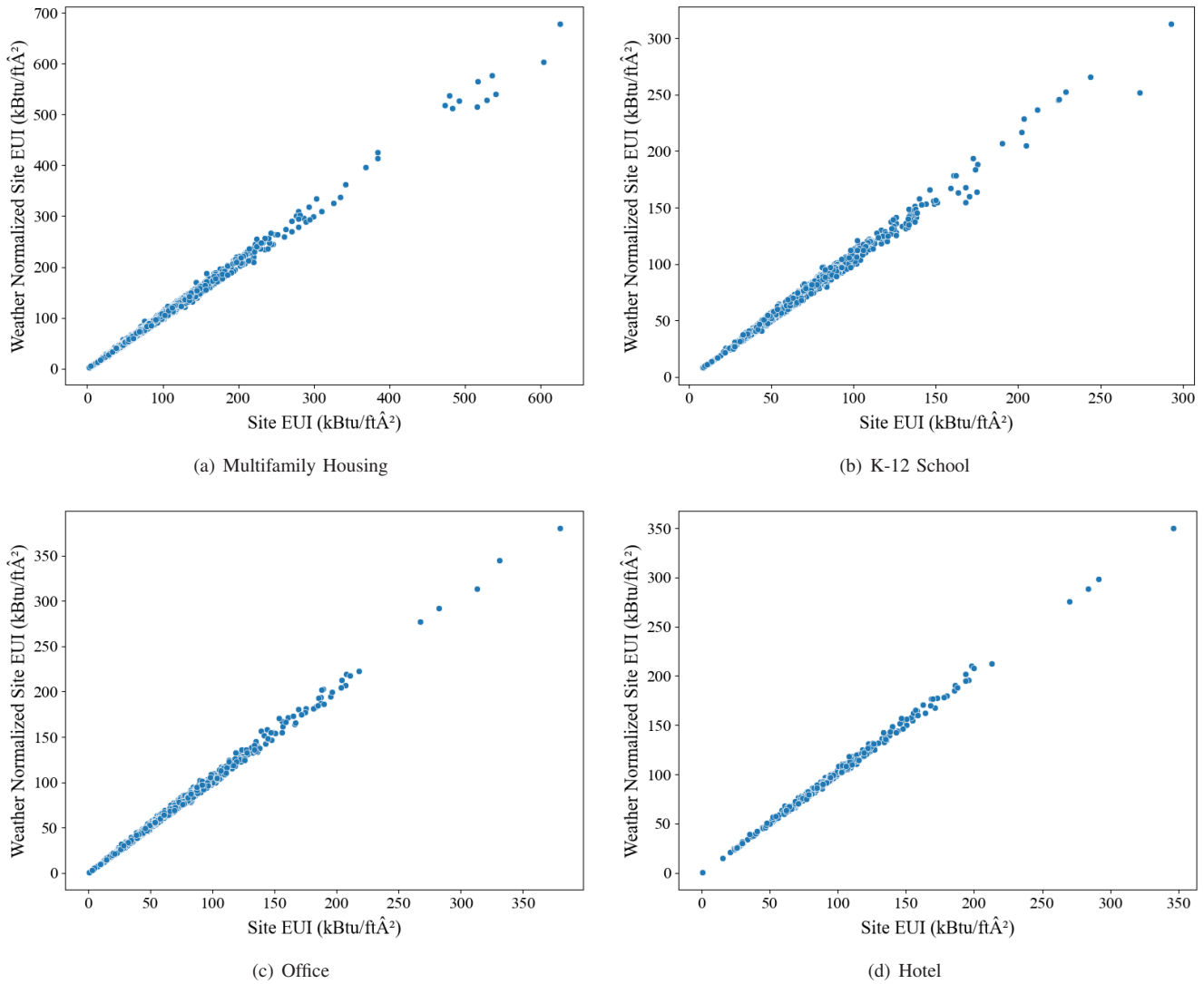


Figure 4. Distribution of correlation between “Site EUI (kBtu/ft²)” and “Weather Normalized Site EUI (kBtu/ft²)”.

consumption of a building or site to its area. Across all categories, an exceptionally strong positive linear relationship is observed, with correlation coefficients approaching 1. This near-perfect correlation is evident in the tight clustering of data points along the diagonal in each scatter plot. After checking the data dictionary, we find that “Site EUI” refers to the site energy use divided by the property square foot; the “Weather Normalized Site EUI” refers to the energy use one property would have consumed during 30-year average weather conditions [40]. Since the “Weather Normalized Site EUI” is calculated based on the “Site EUI”, there is no doubt that there is such a high correlation between these two features. This high multicollinearity suggests that including both variables in predictive models would be redundant and potentially destabilizing. Therefore, only one of the features needs to be retained in the later modeling process. Given its more straightforward interpretation and direct measurement, we opt for keeping the “Site EUI” feature.

B. Feature Selection and Feature Engineering

Due to data measurement and collection challenges, we addressed missing data and potential multicollinearity by implementing a rigorous feature selection process. We removed features with substantial missing data and applied a correlation threshold of 0.7 to filter out highly correlated variables. This careful selection process yielded distinct sets of numeric features for each building type: 7 for multifamily housing, 8 for K-12 schools, 7 for offices, and 5 for hotel buildings. These selected features exhibit correlations below 0.7 with each other, as depicted in Figure 5, ensuring a balanced representation of predictors while minimizing redundancy.

During the feature selection stage, we also engage in feature engineering. Feature engineering entails the extraction or creation of new features from raw data, often involving the transformation of certain raw variables. This may include applying natural logarithm transformations to non-normally

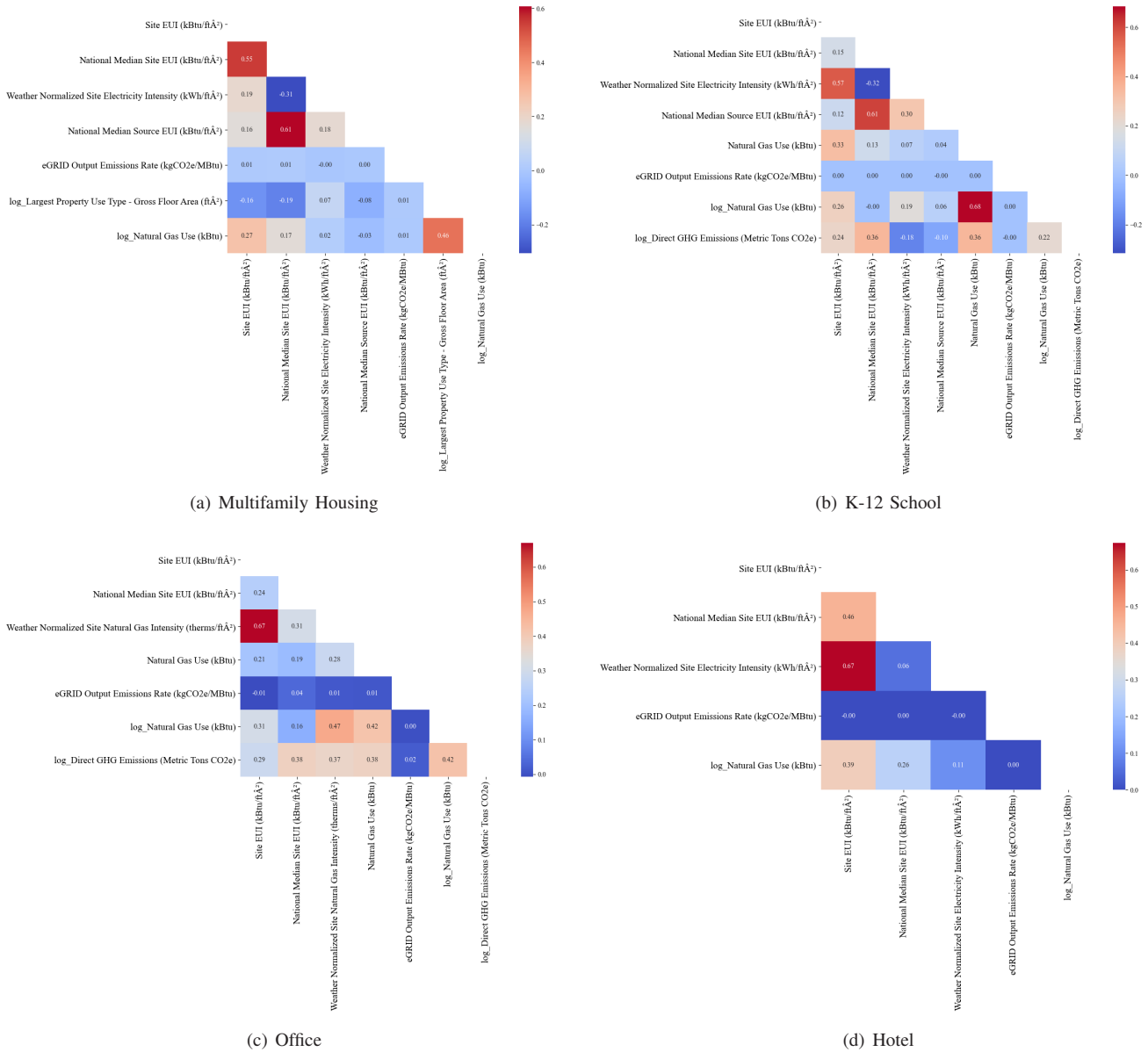


Figure 5. Correlation matrix of selected features.

distributed data or encoding categorical variables with one-hot codes to facilitate their inclusion in model training.

First, we apply the logarithms to the numeric features and add them to the original data. As we all know, most original data are not normally distributed. If we include this kind of data in the model directly, it might arise bias due to the skewed distribution of data. In Figure 5, the features starting with “log_” are the ones transformed by the logarithm functions.

Next, we apply the Min-Max normalization to the numerical features. Scaling these features to a comparable range helps mitigate bias toward features with larger scales, thereby fostering more accurate predictions and enhancing stability. With this step completed, our dataset is now fully prepared for the modeling phase.

C. Test Bench

Our primary objective is to determine the model which best predicts the Energy Star Score of residential buildings. To achieve this goal, we split the dataset into two parts, 70% for training and 30% for testing. We enumerate a combination of different parameters and perform a 4-folds cross-validation to optimize each training model. The training model with the best performance under certain configuration will be used for the testing dataset. The entire experiment is repeated five times, and the average score and standard deviation are reported as the final results. Here, we list the parameters used for each model in the optimization process in Python 3.8.5:

- *k*-Nearest Neighbor Regression:
 - n_neighbors: [5, 10, 15, 20],

TABLE I
SUMMARY OF RESULTS OF THE CASE STUDY.

Building Type	Regressor	MAE	R-squared	Adjusted R-squared	SSE
Multifamily Housing	KNN	12.05±0.70	0.6718±0.0281	0.6655±0.0284	609722.05±52117.79
	Linear	10.96±0.24	0.6528±0.0638	0.6510±0.0642	955467.21±168765.45
	Ridge	11.49±0.33	0.6584±0.0478	0.6566±0.0480	940514.86±124643.94
	DT	1.41±0.05	0.9923±0.0011	0.9923±0.0011	21124.00±2978.79
	RF	2.49±2.07	0.9739±0.0276	0.9722±0.0282	48549.10±51378.32
	SV	6.73±1.82	0.8320±0.0354	0.8288±0.0360	312705.89±68675.44
	GB	0.89±0.08	0.9967±0.0004	0.9966±0.0004	6199.90±806.99
	XGB	1.16±0.04	0.9961±0.0006	0.9961±0.0006	10615.36±1691.00
	MLP	1.16±0.48	0.9937±0.0033	0.9937±0.0033	17512.53±9576.36
K-12 School	KNN	11.62±0.21	0.6450±0.0282	0.6308±0.0293	110224.45±8782.34
	Linear	6.46±0.14	0.8814±0.0125	0.8767±0.0130	36829.02±4025.72
	Ridge	7.16±0.16	0.8705±0.0092	0.8654±0.0095	40199.84±2926.95
	DT	2.75±0.20	0.9683±0.0078	0.9671±0.0082	9814.60±2335.98
	RF	1.82±0.08	0.9891±0.0010	0.9887±0.0010	3385.71±303.33
	SVR	17.31±0.27	0.2683±0.0240	0.2391±0.0249	227165.67±7852.88
	GB	1.44±0.11	0.9909±0.0014	0.9906±0.0014	2812.97±436.87
	XGB	1.76±0.08	0.9889±0.0014	0.9884±0.0014	3452.75±423.18
	MLP	3.40±0.26	0.9615±0.0063	0.9599±0.0065	11958.46±1934.40
Office	KNN	13.17±0.69	0.4940±0.0462	0.4668±0.0487	103329.18±9109.76
	Linear	7.00±0.26	0.7791±0.0587	0.7672±0.0619	45638.35±14317.48
	Ridge	7.97±0.49	0.7781±0.0260	0.7661±0.0274	45655.37±7580.89
	DT	2.99±0.16	0.9630±0.0053	0.9610±0.0056	7574.80±1146.03
	RF	1.69±0.14	0.9854±0.0022	0.9847±0.0023	2984.32±491.18
	SVR	16.65±0.66	0.1861±0.0508	0.1423±0.0535	166443.00±12696.20
	GB	1.72±0.15	0.9894±0.0021	0.9888±0.0022	2169.33±408.58
	XGB	1.90±0.16	0.9852±0.0042	0.9845±0.0044	2998.26±789.87
	MLP	6.99±0.78	0.8110±0.0437	0.8009±0.0461	39047.51±10999.97
Hotel	KNN	18.08±1.45	0.4294±0.0794	0.3455±0.0910	56948.41±9831.27
	Linear	10.18±0.54	0.6932±0.0959	0.6481±0.1100	30186.06±8709.30
	Ridge	12.82±0.96	0.6482±0.0618	0.5965±0.0709	34813.92±5025.98
	DT	6.86±0.31	0.8733±0.0154	0.8547±0.0177	12622.80±1825.87
	RF	4.96±0.46	0.9337±0.0189	0.9239±0.0216	6596.78±1943.27
	SVR	23.65±1.03	0.0729±0.0330	-0.0634±0.0378	92344.71±7476.52
	GB	4.22±0.35	0.9483±0.0162	0.9407±0.0185	5150.55±1687.13
	XGB	4.44±0.33	0.9398±0.0257	0.9309±0.0295	5978.02±2633.96
	MLP	15.62±1.15	0.5626±0.0335	0.4982±0.0384	43513.61±3876.85

- weights: ['uniform', 'distance'],
- algorithm: ['auto', 'ball_tree', 'kd_tree', 'brute'],
- leaf_size: [30, 40, 50]
- Ridge Regression:
 - alpha: [0.1, 1, 10, 100, 1000],
 - solver: ['auto', 'svd', 'cholesky', 'lsqr', 'sparse_cg'],
- Decision Tree Regressor:
 - criterion: ['squared_error', 'absolute_error', 'poisson'],
 - max_depth: [None, 2, 5, 10, 15],
 - min_samples_split: [2, 5, 10, 15],
 - min_samples_leaf: [1, 2, 4, 6],
 - max_features: [None, 'sqrt', 'log2']
- Random Forest Regression:
 - n_estimators: [100, 500, 900, 1100, 1500],
 - max_depth: [None, 2, 5, 10, 15],
 - min_samples_leaf: [1, 2, 4, 6, 8],
 - min_samples_split: [2, 4, 6, 10],
 - max_features: ['sqrt', None, 1]
- Support Vector Regression:
 - C: [0.1, 1, 10, 100],
 - kernel: ['linear', 'poly', 'rbf', 'sigmoid'],
 - gamma: ['scale', 'auto']
- Gradient Boosting Regression:
 - loss: ['squared_error', 'absolute_error', 'huber'],

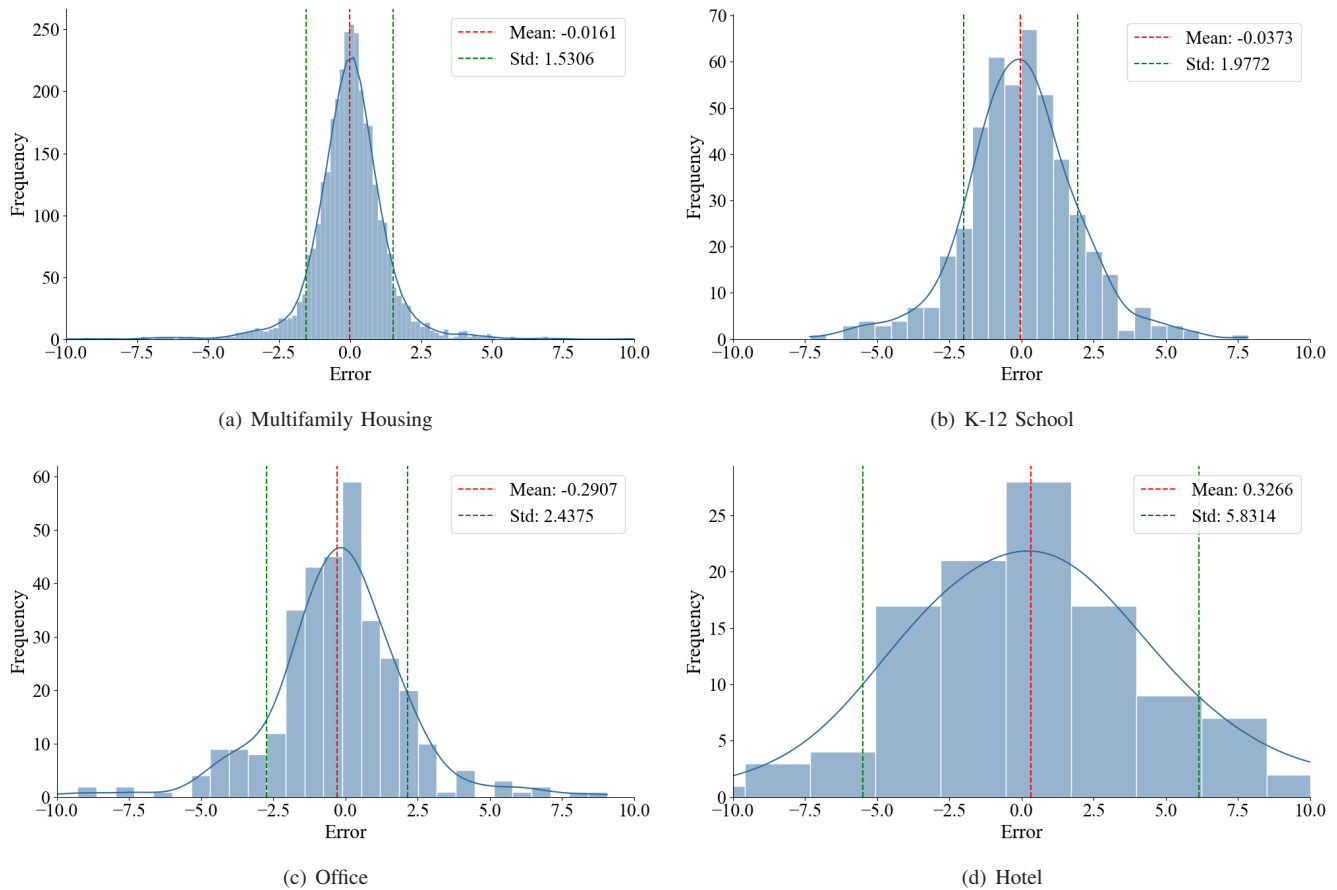


Figure 6. Distribution of residuals.

- n_estimators: [100, 500, 900, 1100, 1500],
- max_depth: [None, 2, 5, 10, 15],
- min_samples_leaf: [1, 2, 4, 6, 8],
- min_samples_split: [2, 4, 6, 10],
- max_features: ['sqrt', None, 1]
- XGBRegressor:
 - n_estimators: [100, 200, 500, 1000],
 - max_depth: [3, 5, 7, 10],
 - learning_rate: [0.01, 0.1, 0.2, 0.3],
 - subsample: [0.8, 0.9, 1.0],
 - colsample_bytree: [0.8, 0.9, 1.0],
 - gamma: [0, 0.1, 0.2, 0.5]
- MLPRegression:
 - hidden_layer_sizes: [(50,), (100,), (100, 50), (100, 100)],
 - activation: ['identity', 'logistic', 'tanh', 'relu'],
 - solver: ['lbfgs', 'sgd', 'adam'],
 - alpha: [0.0001, 0.001, 0.01, 0.1],
 - learning_rate: ['constant', 'invscaling', 'adaptive'],
 - max_iter: [200, 500, 1000]

Note that, there are no hyperparameters in Linear Regression, since its model parameters are determined directly by minimizing the least squares loss function. All machine learning models

were implemented using Python with the Scikit-learn library, and the development environment was PyCharm Community Edition. Scikit-learn is a widely-used, open-source machine learning library that provides simple and efficient tools for data mining and data analysis. Detailed documentation and source code can be found on the official website [41].

D. Results

The analysis of Energy Star Score predictions across four building types in New York City consistently demonstrates the superiority of Gradient Boosting Regression (GBR), which achieves the lowest Mean Absolute Error (MAE) and Sum of Squared Errors (SSE), with R-squared values closest to 1 across all categories. GBR excels in multifamily housing (MAE: 0.89, R-squared: 0.9967), K-12 schools (MAE: 1.44, R-squared: 0.9909), offices (MAE: 1.72, R-squared: 0.9894), and hotels (MAE: 4.22, R-squared: 0.9483). Random Forest (RF) consistently ranks second, performing strongly in K-12 schools (MAE: 1.82, R-squared: 0.9891) and offices (MAE: 1.69, R-squared: 0.9854), while Extreme Gradient Boosting (XGB) follows closely. Support Vector Regression (SVR) shows inconsistent performance, ranging from poor in hotel (MAE: 23.65, R-squared: 0.0729) to moderate in multifamily housing (MAE: 6.73, R-squared: 0.8320). Simpler models like

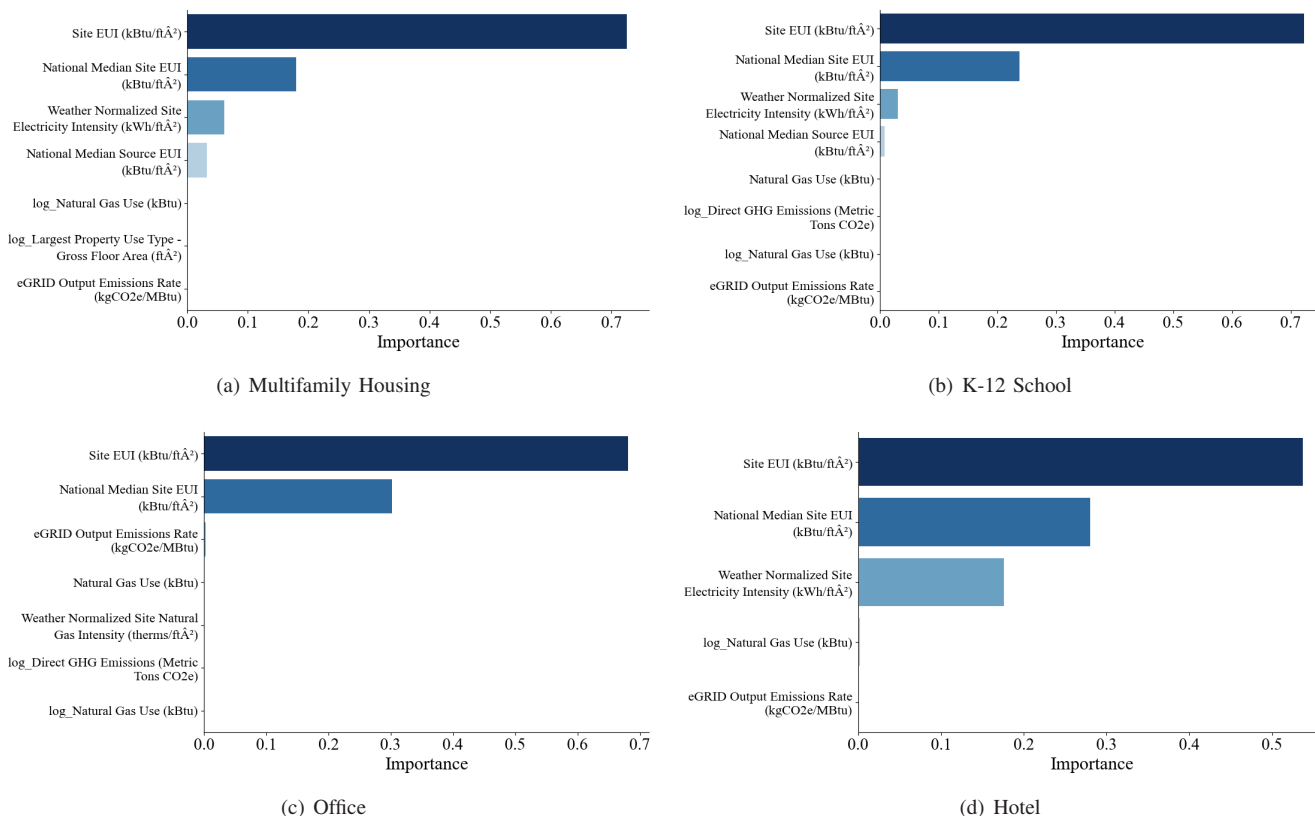


Figure 7. Distribution of importance ranking for the selected features.

KNN, Linear Regression, and Ridge Regression consistently underperform, with KNN showing particularly high MAEs across all types, especially in hotels (MAE: 18.08, R-squared: 0.4294). These results indicate the effectiveness of ensemble and boosting methods in accurately predicting Energy Star Scores for diverse urban building types while highlighting the limitations of simpler models in capturing the complex, non-linear relationships in building energy performance.

Given that the GBR model yielded the best performance across four types of buildings, we analyzed the residual distributions produced by GBR, as shown in Figure 6. Multifamily housing, with the largest dataset of 13,871 samples, shows the best performance with a mean error of -0.0161 and the lowest standard deviation of 1.5306, indicating highly precise and unbiased predictions. K-12 schools follow with a mean error of -0.0373 and a standard deviation of 1.9772. Office buildings show slightly less precision with a mean error of -0.2907 and a standard deviation of 2.4375. Hotels, with the smallest dataset of 393 samples, exhibit the highest variability with a mean error of 0.3266 and a standard deviation of 5.8314. The increasing standard deviations from multifamily housing to hotels directly correspond to the decreasing sample sizes, ranging from 13,871 to 393. Despite the differences in standard deviations, all distributions approximately follow a normal curve centered near zero, indicating that the regression models provide generally reliable predictions across all building types.

Overall, the predictive performance is satisfactory and can offer valuable reference information for decision-makers in energy management and building efficiency across different types.

The feature importance analysis across all four building types reveals consistent patterns with some notable variations. For all building categories, "Site EUI" emerges as the most critical factor, with importance values ranging from approximately 0.5 to 0.7. "National Median Site EUI" consistently ranks second in importance across all types, though its influence varies, being particularly strong for offices and hotels. Hotels demonstrate a unique pattern with "Weather Normalized Site Electricity Intensity" having a notably higher importance, ranking third and showing more significance compared to other building types. For offices, the first two factors, "Site EUI" and "National Median Site EUI", significantly influence the model, with other factors showing much less importance. Multifamily housing shows a more balanced distribution of importance among secondary factors, with "National Median Source EUI" ranking fourth and contributing noticeably to the model. Across all building types, factors related to natural gas use and emissions generally show lower importance, though their rankings vary slightly between categories. This analysis highlights that while energy use intensity metrics are universally crucial for predicting Energy Star Scores, the relative importance of secondary factors can differ based on the specific building type, reflecting the unique energy consumption

patterns and characteristics of each category.

The importance values below 0.01 for the remaining features suggest that they have minimal influence on the model's predictions and can be considered less critical in explaining the variability in the Energy Star Score.

IV. CONCLUSION AND FUTURE WORK

Regression methods have been successfully applied to analyze and model Energy Star Scores across residential, educational, commercial, and lodging structures in New York City. Our comprehensive study, employing nine distinct regression models for these four building types, consistently demonstrates the superiority of the Gradient Boosting Regression (GBR) model. GBR outperforms other methods, achieving the best predictions with minimum errors and variances across all building types. Furthermore, the analysis highlights the universal importance of energy use intensity metrics, particularly "Site EUI" and "National Median Site EUI", while revealing varying influences of secondary factors specific to each building category. Moreover, accurately predicting building energy scores across various types will provide decision-makers with crucial information for retrofitting existing buildings and designing new, energy-efficient structures, ultimately contributing to reduced energy consumption, lower carbon emissions, and more sustainable urban development. Notably, our findings also indicate that the quantity of available data could impact model's stability, with larger datasets for multifamily housing buildings yielding less standard deviations compared to smaller datasets of hotels. Future research will focus on real-time energy emissions analysis and detailed energy usage distribution patterns to further refine energy conservation strategies across various building types.

V. ACKNOWLEDGMENTS

The work is partially supported by the National Science Foundation under NSF Awards Nos. 2234911, 2209637, 2100134. Any opinions, findings, or recommendations, expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

- [1] F. Zhang, B. Chen, F. Wu, and L. Bai, "Prediction of residential building energy star score: A case study of new york city", in *2024 IARIA Annual Congress on Frontiers in Science, Technology, Services, and Applications*, Porto, Portugal, Jul. 2024, pp. 180–186, ISBN: 978-1-68558-180-0.
- [2] J. Syvitski *et al.*, "Extraordinary human energy consumption and resultant geological impacts beginning around 1950 ce initiated the proposed anthropocene epoch", *Communications Earth & Environment*, vol. 1, no. 1, p. 32, 2020.
- [3] P. Nejat, F. Jomehzadeh, M. M. Taheri, M. Gohari, and M. Z. A. Majid, "A global review of energy consumption, co2 emissions and policy in the residential sector (with an overview of the top ten co2 emitting countries)", *Renewable and Sustainable Energy Reviews*, vol. 43, pp. 843–862, 2015.
- [4] K. Binita and M. Ruth, "Estimation and projection of institutional building electricity consumption", *Energy and Buildings*, vol. 143, pp. 43–52, 2017.
- [5] W. Feist and J. Schnieders, "Energy efficiency—a key to sustainable housing", *The European Physical Journal Special Topics*, vol. 176, no. 1, pp. 141–153, 2009.
- [6] *An assessment of energy technologies and research opportunities*, https://www.energy.gov/sites/prod/files/2015/09/f26/Quadrennial-Technology-Review-2015_0.pdf, Accessed: Aug 22, 2024, 2015.
- [7] M. Santamouris and K. Vasilakopoulou, "Present and future energy consumption of buildings: Challenges and opportunities towards decarbonisation", *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, vol. 1, p. 100002, 2021.
- [8] *Our history about energy star*, <https://www.energystar.gov/about/how-energy-star-works/history>, Accessed: Aug 11, 2024, 2024.
- [9] T. W. Hicks and B. Von Neida, "US national energy performance rating system and energy star building certification program", in *Proceedings of the 2004 Improving Energy Efficiency of Commercial Buildings Conference*, 2004, pp. 1–9.
- [10] D. B. Crawley, "Building energy tools directory", *Proceedings of Building Simulation'97*, vol. 1, pp. 63–64, 1997.
- [11] K. Amasyali and N. M. El-Gohary, "A review of data-driven building energy consumption prediction studies", *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1192–1205, 2018.
- [12] S. Fathi, R. Srinivasan, A. Fenner, and S. Fathi, "Machine learning applications in urban building energy performance forecasting: A systematic review", *Renewable and Sustainable Energy Reviews*, vol. 133, p. 110287, 2020.
- [13] Q. Qiao, A. Yunusa-Kaltungo, and R. E. Edwards, "Towards developing a systematic knowledge trend for building energy consumption prediction", *Journal of Building Engineering*, vol. 35, p. 101967, 2021.
- [14] T. Ahmad *et al.*, "Supervised based machine learning models for short, medium and long-term energy prediction in distinct building environment", *Energy*, vol. 158, pp. 17–32, 2018.
- [15] H. C. Jung, J. S. Kim, and H. Heo, "Prediction of building energy consumption using an improved real coded genetic algorithm based least squares support vector machine approach", *Energy and Buildings*, vol. 90, pp. 76–84, 2015.
- [16] Z. Ma, C. Ye, and W. Ma, "Support vector regression for predicting building energy consumption in southern china", *Energy Procedia*, vol. 158, pp. 3433–3438, 2019.

- [17] Z. Yu, F. Haghghat, B. C. Fung, and H. Yoshino, "A decision tree method for building energy demand modeling", *Energy and Buildings*, vol. 42, no. 10, pp. 1637–1646, 2010.
- [18] Y. Liu, H. Chen, L. Zhang, and Z. Feng, "Enhancing building energy efficiency using a random forest model: A hybrid prediction approach", *Energy Reports*, vol. 7, pp. 5003–5012, 2021.
- [19] S. Touzani, J. Granderson, and S. Fernandes, "Gradient boosting machine for modeling the energy consumption of commercial buildings", *Energy and Buildings*, vol. 158, pp. 1533–1543, 2018.
- [20] H. Lu, F. Cheng, X. Ma, and G. Hu, "Short-term prediction of building energy consumption employing an improved extreme gradient boosting model: A case study of an intake tower", *Energy*, vol. 203, p. 117756, 2020.
- [21] G. S. Georgiou, P. Christodoulides, and S. A. Kalogirou, "Implementing artificial neural networks in energy building applications—a review", in *2018 IEEE International Energy Conference (ENERGYCON)*, IEEE, 2018, pp. 1–6.
- [22] N. Somu, G. R. MR, and K. Ramamritham, "A deep learning framework for building energy consumption forecast", *Renewable and Sustainable Energy Reviews*, vol. 137, p. 110591, 2021.
- [23] S. Afzal, B. M. Ziapour, A. Shokri, H. Shakibi, and B. Sobhani, "Building energy consumption prediction using multilayer perceptron neural network-assisted models; comparison of different optimization algorithms", *Energy*, vol. 282, p. 128446, 2023.
- [24] F. Wahid, D. Kim, *et al.*, "A prediction approach for demand analysis of energy consumption using k-nearest neighbor in residential buildings", *International Journal of Smart Home*, vol. 10, no. 2, pp. 97–108, 2016.
- [25] N. Fumo and M. R. Biswas, "Regression analysis for prediction of residential energy consumption", *Renewable and Sustainable Energy Reviews*, vol. 47, pp. 332–343, 2015.
- [26] E. K. Laitinen and T. Laitinen, "Bankruptcy prediction: Application of the Taylor's expansion in logistic regression", *International Review of Financial Analysis*, vol. 9, no. 4, pp. 327–349, 2000.
- [27] D. J. Briggs *et al.*, "A regression-based method for mapping traffic-related air pollution: Application and testing in four contrasting urban environments", *Science of the Total Environment*, vol. 253, no. 1-3, pp. 151–167, 2000.
- [28] E. Suárez, C. M. Pérez, R. Rivera, and M. N. Martínez, *Applications of Regression Models in Epidemiology*. John Wiley & Sons, 2017.
- [29] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression", *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.
- [30] J. Groß, *Linear regression*. Springer Science & Business Media, 2003, vol. 175.
- [31] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems", *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.
- [32] W.-Y. Loh, "Classification and regression trees", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 14–23, 2011.
- [33] M. R. Segal, "Machine learning benchmarks and random forest regression", *UCSF: Center for Bioinformatics and Molecular Biostatistics*, 2004.
- [34] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines", *Advances in Neural Information Processing Systems*, vol. 9, pp. 155–161, 1996.
- [35] N. Duffy and D. Helmbold, "Boosting methods for regression", *Machine Learning*, vol. 47, pp. 153–200, 2002.
- [36] 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.
- [37] M. W. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences", *Atmospheric Environment*, vol. 32, no. 14-15, pp. 2627–2636, 1998.
- [38] V. Plevris, G. Solorzano, N. P. Bakas, and M. E. A. Ben Seghier, "Investigation of performance metrics in regression analysis and machine learning-based prediction models", in *8th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS Congress 2022)*, European Community on Computational Methods in Applied Sciences, 2022, pp. 1–25.
- [39] A. V. Tatachar, "Comparative assessment of regression models based on model evaluation metrics", *International Research Journal of Engineering and Technology (IRJET)*, vol. 8, no. 09, pp. 2395–0056, 2021.
- [40] *Energy and Water Data Disclosure for Local Law 84 2022 (Data for Calendar Year 2021)*, <https://data.cityofnewyork.us/Environment/Energy-and-Water-Data-Disclosure-for-Local-Law-84-/7x5e-2fxh>, [Online; retrieved: May,2024].
- [41] *Scikit-learn*, <https://scikit-learn.org>, Accessed: June 22, 2024.

Symbolic Unfolding versus Tuning of Similarity-based Fuzzy Logic Programs

Ginés Moreno

Department of Computing Systems
University of Castilla-La Mancha
02071 Albacete (Spain)
Email: Gines.Moreno@uclm.es

José Antonio Riaza

Department of Computing Systems
University of Castilla-La Mancha
02071 Albacete (Spain)
Email: JoseAntonio.Riaza@uclm.es

Abstract—FASILL introduces “*Fuzzy Aggregators and Similarity Into a Logic Language*”. In its symbolic extension, called sFASILL, some truth degrees, similarity annotations and fuzzy connectives can be left unknown, so that the user can easily figure out the impact of their possible values at execution time. In this paper, we firstly adapt to this last setting a similarity-based, symbolic variant of unfolding rule (very well known in most declarative frameworks), which is based on the application of computational steps on the bodies of program rules for improving efficiency. Next, we combine it with previous tuning techniques intended to transform a symbolic sFASILL program into the concrete customized FASILL one that best satisfies the user’s preferences. The improved methods have been implemented in a freely available online tool, which has served us to develop several experiments and benchmarks evidencing the good performance of the resulting system. To the best of our knowledge, our analysis is the first one combining unfolding and tuning techniques in a fully integrated fuzzy logic programming setting.

Index Terms—Fuzzy Logic Programming; Similarity; Symbolic Unfolding; Tuning.

I. INTRODUCTION

This paper extends our initial approach described in [1] (presented at the 2024 IARIA Annual Congress on *Frontiers in Science, Technology, Services, and Applications – IARIA Congress 2024*), by combining and enriching the fuzzy capabilities of a highly flexible programming environment developed in our research group. During the last four decades, the research field of *Fuzzy Logic Programming* has promoted the introduction of *Fuzzy Logic* [2] concepts into *Logic Programming* [3] in order to deal with vagueness in a natural way [4]. It has provided an extensive variety of logic programming dialects promoting the development of flexible real-world applications in the fields of artificial intelligence, soft-computing, semantic web, etc. Some interesting approaches focus on replacing the classic (syntactic) unification algorithm of Prolog by one based on the use of similarity/proximity relations [5][6], such as Likelog [7] and Bousi~Prolog [8]. Similarity/proximity relations connect the elements of a set with a certain approximation degree and serve for weakening the notion of equality and, hence, to deal with vague information [9]. Other approaches modify the operational principle of pure logic programming to replace it by inference mechanisms based on fuzzy logic, which allow a wide variety of connectives and the use of a gradation of truth degrees (beyond the traditional values of *true* and *false*). Most of these systems implement the fuzzy resolution principle introduced by Lee in [10], such as Prolog-Elf [11], F-Prolog [12],

generalized annotated logic programming [13], (S-)QLP [15], Fril [14], Fuzzy-Prolog [16], RFuzzy [17] and MALP [18].

Since the logic language Prolog has been fuzzified by embedding similarity relations or using fuzzy connectives for dealing with truth degrees beyond $\{true, false\}$, respectively, we have recently combined both approaches in the design of FASILL [19], whose symbolic extension (inspired by our initial experiences with MALP [20]) is called sFASILL [21]. This last symbolic language is useful for *flexibly tuning* (according to users preferences) the fuzzy components of fuzzy logic programs. A tuning problem is a pair composed by a symbolic program plus a set of test cases where users express their wishes about the expected behaviour of the final, customized program. As a very simple example, consider just one symbolic program rule like $p \leftarrow @_{aver}(\#sc, 0.8)$, where $@_{aver}$ refers to the average connective and $\#sc$ is a *symbolic constant* representing an unknown truth degree, together with only a test case of the form $0.6 \rightarrow p$, where a user indicates that (s)he wants to obtain 0.6 when evaluating p . Although this tuning problem has the trivial solution of replacing $\#sc$ by 0.4, in the general case, it is not easy to reach good (usually approximate) solutions when the number of program rules, symbolic constants and test cases grow more and more.

In this paper, we will collect from [21][22] two tuning strategies showing that by “partially” executing symbolic sFASILL programs and then replacing the unknown values and connectives (on their program rules and associated similarity relations) by concrete ones, gives the same result than replacing these values and connectives in the original sFASILL program and, then, fully executing the resulting FASILL program. So, sFASILL programs can be used to automatically tune and synthesize a FASILL program w.r.t. a given set of test cases, thus easing what is considered the most difficult part of the process: the specification of the truth/similarity degrees and connectives in the program. Although there exist other approaches, which are able to *tune* fuzzy truth degrees and connectives [23][24][25], none of them manage similarity relations as the tuning technique we describe in [21] does. Let us mention that we have used sFASILL and its tuning engine for developing two real world applications in the fields of the semantic web [26] and neural networks [27].

Besides this, unfolding is a well-known and widely used semantics-preserving program transformation rule, which is able to improve programs, generating more efficient code. The unfolding transformation traditionally considered in pure logic

programming consists in the replacement of a program clause C by the set of clauses obtained after applying a computation step in all its possible forms on the body of C [28][29].

In order to briefly illustrate the essence and benefits of the transformation, consider a very simple Prolog program containing a clause, say $p(X):-q(X)$, and a fact, say $q(a)$, for defining two (crisp, not fuzzy) predicates, p and q . It is easy to see that both rules must be used in two computational steps for successfully executing a goal like $p(a)$. Alternatively, we can unfold the first clause by applying a computational step on its body $q(X)$ (using the fact $q(a)$) and next instantiating the head with the achieved substitution $\{X/a\}$. Then, the new unfolded rule is just the simple fact $p(a)$, which must be used in only one computational step (instead of two, as before) to solve goal $p(a)$. This very simple example reveals that all computational steps applied at unfolding time *remain compiled* on unfolded rules forever, and hence, those steps have no longer to be repeated in all subsequent executions of the transformed programs. This justifies why unfolding is able to improve the efficiency of transformed programs by accelerating their computational behaviour.

In [30][31], we successfully adapted such operation to fuzzy logic programs dealing with lattices of truth degrees and similarity relations, but this type of unfolding was not symbolic yet. On the contrary, in [32][33] we defined a symbolic version of the transformation but in absence of similarities. Inspired by both works, in [1] we have recently fused both approaches in the definition of a similarity-based symbolic transformation. Now, we extend such work by using this transformation as a pre-process of our tuning engines in order to improve the performance of the resulting system. In this paper we describe and use a freely available online tool ([34]) for developing some revealing experiments, benchmarks and analysis of our approach.

The structure of this paper is as follows. After summarizing, in Section II, the syntax of FASILL and sFASILL, in Section III we detail how to execute and unfold such programs. Next, Section IV summarizes two tuning engines, which are combined with unfolding in Section V, also analyzing its performance and practicability. Finally, we conclude and propose future work in Section VI.

II. THE FASILL LANGUAGE AND ITS SYMBOLIC EXTENSION

In this work, given a complete lattice L , we consider a first order language \mathcal{L}_L built upon a signature Σ_L , that contains the elements of a countably infinite set of variables \mathcal{V} , function and predicate symbols (denoted by \mathcal{F} and Π , respectively) with an associated arity—usually expressed as pairs f/n or p/n , respectively, where n represents its arity—, and the truth degree literals Σ_L^T and connectives Σ_L^C from L . Therefore, a well-formed formula in \mathcal{L}_L can be either:

- A *value* $v \in \Sigma_L^T$, which will be interpreted as itself, i.e., as the truth degree $v \in L$.

- $p(t_1, \dots, t_n)$, if t_1, \dots, t_n are terms over $\mathcal{V} \cup \mathcal{F}$ and p/n is an n -ary predicate. This formula is called *atomic* (atom, for short).
- $\varsigma(e_1, \dots, e_n)$, if e_1, \dots, e_n are well-formed formulas and ς is an n -ary connective with truth function $\llbracket \varsigma \rrbracket : L^n \mapsto L$.

Definition 1 (Complete Lattice). A *complete lattice* is a partially ordered set (L, \leq) such that every subset S of L has infimum and supremum elements. Then, it is a bounded lattice, i.e., it has bottom and top elements, denoted by \perp and \top , respectively.

Example 1. In this paper, we use the lattice $([0, 1], \leq)$, where \leq is the usual ordering relation on real numbers, and three sets of conjunctions/disjunctions corresponding to the fuzzy logics of Gödel, Łukasiewicz and Product (with different capabilities for modelling *pessimistic*, *optimistic* and *realistic scenarios*), defined in Figure 1. It is possible to also include other fuzzy connectives (aggregators) like the arithmetical and geometrical averages, say $@_{\text{aver}}(x, y) \triangleq (x+y)/2$ and $@_{\text{geom}}(x, y) \triangleq \sqrt{xy}$, or the linguistic modifier $@_{\text{very}}(x) \triangleq x^2$.

Definition 2 (Similarity Relation). Given a domain \mathcal{U} and a lattice L with a fixed t-norm \wedge , a *similarity relation* \mathcal{R} is a fuzzy binary relation on \mathcal{U} , that is, a fuzzy subset on $\mathcal{U} \times \mathcal{U}$ (namely, a mapping $\mathcal{R} : \mathcal{U} \times \mathcal{U} \rightarrow L$) fulfilling the following properties: reflexive $\forall x \in \mathcal{U}, \mathcal{R}(x, x) = \top$, symmetric $\forall x, y \in \mathcal{U}, \mathcal{R}(x, y) = \mathcal{R}(y, x)$, and transitive $\forall x, y, z \in \mathcal{U}, \mathcal{R}(x, z) \geq \mathcal{R}(x, y) \wedge \mathcal{R}(y, z)$.

The fuzzy logic language FASILL relies on complete lattices and similarity relations [19]. We are now ready for summarizing its *symbolic* extension where, in essence, we allow some undefined values (truth degrees) and connectives in program rules as well as in the associated similarity relation, so that these elements can be systematically computed afterwards. The symbolic extension of FASILL we initially presented in [21] is called sFASILL.

Given a complete lattice L , we consider an augmented signature $\Sigma_L^\#$ producing an augmented language $\mathcal{L}_L^\# \supseteq \mathcal{L}_L$, which may also include a number of symbolic values and symbolic connectives, which do not belong to L . Symbolic objects are usually denoted as $o^\#$ with a superscript $\#$ and, in our tool, their identifiers always start with $\#$. An $L^\#$ -*expression* is now a well-formed formula of $\mathcal{L}_L^\#$, which is composed by values and connectives from L as well as by symbolic values and connectives. We let $\text{exp}_L^\#$ denote the set of all $L^\#$ -expressions in $\mathcal{L}_L^\#$. Given a $L^\#$ -expression E , $\llbracket E \rrbracket$ refers to the new $L^\#$ -expression obtained after evaluating as much as possible the connectives in E . Particularly, if E does not contain any symbolic value or connective, then $\llbracket E \rrbracket = v \in L$.

In the following, we consider *symbolic substitutions* that are mappings from symbolic values and connectives to expressions over $\Sigma_L^T \cup \Sigma_L^C$. We let $\text{sym}(o^\#)$ denote the symbolic values and connectives in $o^\#$. Given a symbolic substitution Θ for $\text{sym}(o^\#)$, we denote by $o^\# \Theta$ the object that results from $o^\#$ by replacing every symbolic symbol $e^\#$ by $e^\# \Theta$.

$\&_{\text{prod}}(x, y) \triangleq x * y$	$ _{\text{prod}}(x, y) \triangleq x + y - xy$	Product logic
$\&_{\text{godel}}(x, y) \triangleq \min(x, y)$	$ _{\text{godel}}(x, y) \triangleq \max(x, y)$	Gödel logic
$\&_{\text{luka}}(x, y) \triangleq \max(0, x + y - 1)$	$ _{\text{luka}}(x, y) \triangleq \min(x + y, 1)$	Lukasiewicz logic

Fig. 1. Conjunctions and disjunctions of three different fuzzy logics over $([0, 1], \leq)$.

Definition 3 (Symbolic Similarity Relation). Given a domain \mathcal{U} and a lattice L with a fixed —possibly symbolic— t-norm \wedge , a *symbolic similarity relation* is a mapping $\mathcal{R}^\# : \mathcal{U} \times \mathcal{U} \rightarrow \exp_L^\#$ such that, for any symbolic substitution Θ for $\text{sym}(\mathcal{R}^\#)$, the result of fully evaluating all L -expressions in $\mathcal{R}^\#\Theta$, say $\llbracket \mathcal{R}^\#\Theta \rrbracket$, is a similarity relation.

Definition 4 (Symbolic Rule and Symbolic Program). Let L be a complete lattice. A *symbolic rule* over L is a formula $A \leftarrow \mathcal{B}$, where the following conditions hold:

- A is an atomic formula of \mathcal{L}_L (the head of the rule);
- \leftarrow is an implication from L or a symbolic implication;
- \mathcal{B} (the body of the rule) is a symbolic goal, i.e., a well-formed formula of $\mathcal{L}_L^\#$;

A *sFASILL program* is a tuple $\mathcal{P}^\# = \langle \Pi^\#, \mathcal{R}^\#, L \rangle$ where $\Pi^\#$ is a set of symbolic rules, $\mathcal{R}^\#$ is a symbolic similarity relation between the elements of the signature Σ of $\Pi^\#$, and L is a complete lattice.

Example 2. Consider a symbolic sFASILL program $\mathcal{P}^\# = \langle \Pi^\#, \mathcal{R}^\#, L \rangle$ based on lattice $L = ([0, 1], \leq)$, where $\Pi^\#$ is the following set of symbolic rules:

$$\Pi^\# = \left\{ \begin{array}{l} R_1 : \text{vanguardist}(\text{ritz}) \leftarrow 0.9 \\ R_2 : \text{elegant}(\text{hydropolis}) \leftarrow s_3^\# \\ R_3 : \text{close}(\text{hydropolis}, \text{taxi}) \leftarrow 0.7 \\ R_4 : \text{good_hotel}(x) \leftarrow \\ \quad @_{s_4}^\#(\text{elegant}(x), @_{\text{very}}(\text{close}(x, \text{metro}))) \end{array} \right.$$

Note here that we leave unknown the level in which the hotel *hydropolis* is more or less elegant (see the symbolic constant $s_3^\#$ in the second fact) as well as which should be the most appropriate connective for combining two features required on good hotels (see the symbolic constant $@_{s_4}^\#$ in the body of the fourth rule).

The symbolic similarity relation $\mathcal{R}^\#$ on $\mathcal{U} = \{\text{vanguardist}, \text{elegant}, \text{modern}, \text{metro}, \text{taxi}, \text{bus}\}$, is represented by the graph shown in Figure 2 (a matrix can be also used to represent this concept).

This symbolic similarity relation $\mathcal{R}^\#$ has been obtained after applying the closure algorithm we initially introduced in [21], which is inspired by [35][36][37] and, in essence, is an adaptation of the classical Warshall's algorithm for computing transitive closures. In this particular example, we have selected the symbolic t-norm $\&_{s_2}^\#$ and the following set of similarity equations: $\text{elegant} \sim \text{modern} = s_0^\#$, $\text{modern} \sim \text{vanguardist} = 0.9$, $\text{metro} \sim \text{bus} = 0.5$ and $\text{bus} \sim \text{taxi} = s_1^\#$.

In what follows, we plan to introduce and combine the unfolding and tuning techniques we have developed in the last years for reinforcing their power in this novel, symbolic plus similarity-based fuzzy logic setting.

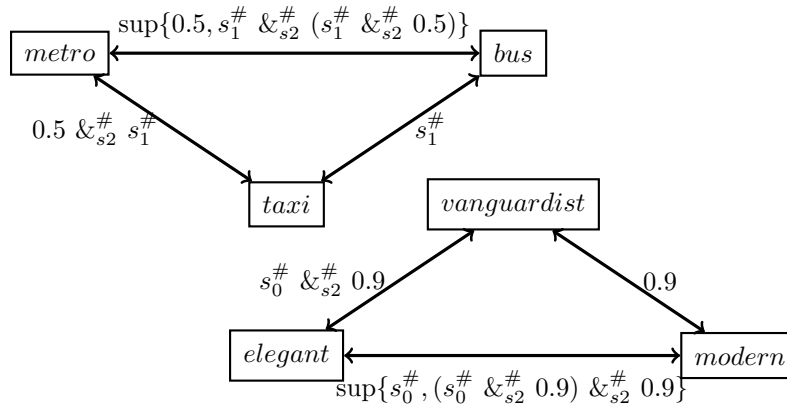
III. RUNNING AND UNFOLDING sFASILL PROGRAMS

As a logic language, sFASILL inherits the concepts of substitution, unifier and most general unifier (*mgu*) from pure logic programming, but extending some of them in order to cope with similarities, as Bousi~Prolog [8] does, where the concept of most general unifier is replaced by the one of *weak most general unifier* (w.m.g.u.). One step beyond, in [21] we extended again this notion by referring to *symbolic weak most general unifiers* (s.w.m.g.u.) and a symbolic weak unification algorithm was introduced to compute them. Roughly speaking, the *symbolic weak unification algorithm* states that two *expressions* (i.e., terms or atomic formulas) $f(t_1, \dots, t_n)$ and $g(s_1, \dots, s_n)$ weakly unify if the root symbols f and g are close with a certain —possibly symbolic— degree (i.e., $\mathcal{R}^\#(f, g) = r \neq \perp$) and each of their arguments t_i and s_i weakly unify. Therefore, there is a symbolic weak unifier for two expressions even if the symbols at their roots are not syntactically equal ($f \neq g$).

More technically, the symbolic weak unification algorithm can be seen as an reformulation/extension of the ones appearing in [6] (since now we manage arbitrary complete lattices) and [19][8] (because now we deal with symbolic similarity relations). In essence, the *symbolic weak most general unifier* of two expressions \mathcal{E}_1 and \mathcal{E}_2 , say $\text{wmgu}^\#(\mathcal{E}_1, \mathcal{E}_2) = \langle \sigma, E \rangle$, is the simplest *symbolic substitution* σ of \mathcal{E}_1 and \mathcal{E}_2 together with its *symbolic unification degree* E verifying that $E = \hat{\mathcal{R}}(E_1\sigma, E_2\sigma)$.

Example 3. Given the complete lattice $L = ([0, 1], \leq)$ of Example 1 and the symbolic similarity relation $\mathcal{R}^\#$ of Example 2, we can use the symbolic t-norm $\&_{s_2}^\#$ for computing the following two symbolic symbolic weak most general unifiers: $\text{wmgu}^\#(\text{modern}(\text{taxi}), \text{vanguardist}(\text{bus})) = \langle \{\}, 0.9 \&_{s_2}^\# s_1^\# \rangle$ and $\text{wmgu}^\#(\text{close_to}(X, \text{taxi}), \text{close_to}(\text{ritz}, \text{bus})) = \langle \{X/\text{ritz}\}, s_1^\# \rangle$

In order to describe the procedural semantics of the sFASILL language, in the following, we denote by $\mathcal{C}[A]$ a formula where A is a sub-expression (usually an atom), which occurs in the —possibly empty— context $\mathcal{C}[\]$ whereas $\mathcal{C}[A/A']$ means the replacement of A by A' in the context $\mathcal{C}[\]$. Moreover, $\text{Var}(s)$ denotes the set of distinct variables occurring in the syntactic object s and $\theta[\text{Var}(s)]$ refers to the substitution obtained from θ by restricting its domain to $\text{Var}(s)$. In the next definition, we always consider that A is the


 Fig. 2. Example of symbolic similarity relation $\mathcal{R}^\#$.

selected atom in a goal Q , L is the complete lattice associated to $\Pi^\#$ and, as usual, rules are renamed apart:

Definition 5 (Computational Step). Let Q be a goal and σ a substitution. The pair $\langle Q; \sigma \rangle$ is a *state*. Given a symbolic program $\langle \Pi^\#, \mathcal{R}^\#, L \rangle$ and a (possibly symbolic) t-norm \wedge in L , a *computation* is formalized as a state transition system, whose transition relation \rightsquigarrow is the smallest relation satisfying these rules:

1) *Successful step* (denoted as $\overset{SS}{\rightsquigarrow}$):

$$\frac{A' \leftarrow B \in \Pi^\# \quad \langle Q[A], \sigma \rangle \quad \text{wmg}u^\#(A, A') = \langle \theta, E \rangle \quad E \neq \perp}{\langle Q[A/E \wedge B]\theta, \sigma\theta \rangle} \text{SS}$$

2) *Failure step* (denoted as $\overset{FS}{\rightsquigarrow}$):

$$\frac{\langle Q[A], \sigma \rangle \quad \nexists A' \leftarrow B \in \Pi^\# : \text{wmg}u^\#(A, A') = \langle \theta, E \rangle}{\langle Q[A/\perp], \sigma \rangle} \text{FS}$$

3) *Interpretive step* (denoted as $\overset{IS}{\rightsquigarrow}$):

$$\frac{\langle Q; \sigma \rangle \text{ where } Q \text{ is a } L^\# \text{-expression}}{\langle \llbracket Q \rrbracket; \sigma \rangle} \text{IS}$$

Definition 6 (Derivation and Symbolic Fuzzy Computed Answer). A *derivation* is a sequence of arbitrary length $\langle Q; id \rangle \rightsquigarrow^* \langle Q'; \sigma \rangle$. When Q' is an $L^\#$ -expression that cannot be further reduced, $\langle Q'; \sigma' \rangle$, where $\sigma' = \sigma[\text{Var}(Q)]$, is called a *symbolic fuzzy computed answer* (sfca). Also, if Q' is a concrete value of L , we say that $\langle Q'; \sigma' \rangle$ is a *fuzzy computed answer* (fca).

The following example illustrates the operational semantics of sFASILL.

Example 4. Let $\mathcal{P}^\# = \langle \Pi^\#, \mathcal{R}^\#, L \rangle$ be the program from Example 2. It is possible to perform the following derivation for $\mathcal{P}^\#$ and goal $Q = \text{good_hotel}(x)$ obtaining the sfca

$$\langle Q_1; \sigma_1 \rangle = \langle @_{s_4}^\#(\&_{s_2}^\#(\&_{s_2}^\#(s_0^\#, 0.9), 0.9), 0.0); \{x/\text{ritz}\} \rangle:$$

$$\langle \text{good_hotel}(x), id \rangle \overset{SS}{\rightsquigarrow} \text{R4}$$

$$\langle @_{s_4}^\#(\text{elegant}(x_1), @_{\text{very}}(\text{close}(x_1, \text{metro}))), \{x/x_1\} \rangle \overset{SS}{\rightsquigarrow} \text{R1}$$

$$\langle @_{s_4}^\#(\&_{s_2}^\#(\&_{s_2}^\#(s_0^\#, 0.9), 0.9), @_{\text{very}}(\text{close}(\text{ritz}, \text{metro}))), \{x/\text{ritz}\} \rangle \overset{FS}{\rightsquigarrow}$$

$$\langle @_{s_4}^\#(\&_{s_2}^\#(\&_{s_2}^\#(s_0^\#, 0.9), 0.9), @_{\text{very}}(0.0)), \{x/\text{ritz}\} \rangle \overset{IS}{\rightsquigarrow}$$

$$\langle @_{s_4}^\#(\&_{s_2}^\#(\&_{s_2}^\#(s_0^\#, 0.9), 0.9), 0.0), \{x/\text{ritz}\} \rangle$$

Apart from this derivation, there exists a second one ending with the alternative sfca $\langle Q_2; \sigma_2 \rangle = \langle @_{s_4}^\#(s_3^\#, @_{\text{very}}(\&_{s_2}^\#(\&_{s_2}^\#(0.5, s_1^\#), 0.7))), \{x/\text{hydropolis}\} \rangle$ associated to the same goal. Both derivations are included in the tree shown in Figure 3. Observe the presence of symbolic constants coming from the symbolic similarity relation, which contrast with our precedent work [20].

Now, let $\Theta = \{s_0^\#/0.8, s_1^\#/0.8, \&_{s_2}^\#/\&_{1\text{uka}}, s_3^\#/1.0, @_{s_4}^\#/@_{\text{aver}}\}$ be a symbolic substitution that can be used for instantiating the previous sFASILL program in order to obtain a non-symbolic, fully executable FASILL program. This substitution can be automatically obtained by the tuning engines we will describe in Section IV ([21]) after introducing a couple of test cases (i.e., $0.4 \rightarrow \text{good_hotel}(\text{hydropolis})$ and $0.6 \rightarrow \text{good_hotel}(\text{ritz})$), which represent the desired degrees for two goals accordingly to the user preferences.

Now we are ready to introduce the similarity-based symbolic unfolding transformations relying on the operational semantics described so far.

Definition 7 (Symbolic Unfolding). Let $\mathcal{P}^\# = \langle \Pi^\#, \mathcal{R}^\#, L \rangle$ be a sFASILL program and $R : (H \leftarrow B) \in \Pi^\#$ be a rule (with non-empty body B). Then, the *symbolic unfolding* of rule R in program $\mathcal{P}^\#$ is the new sFASILL program $\mathcal{P}'^\# = \langle \Pi'^\#, \mathcal{R}^\#, L \rangle$, where $\Pi'^\# = (\Pi^\# - \{R\}) \cup \{H\sigma \leftarrow B' \mid \langle B; id \rangle \rightsquigarrow \langle B'; \sigma \rangle\}$.

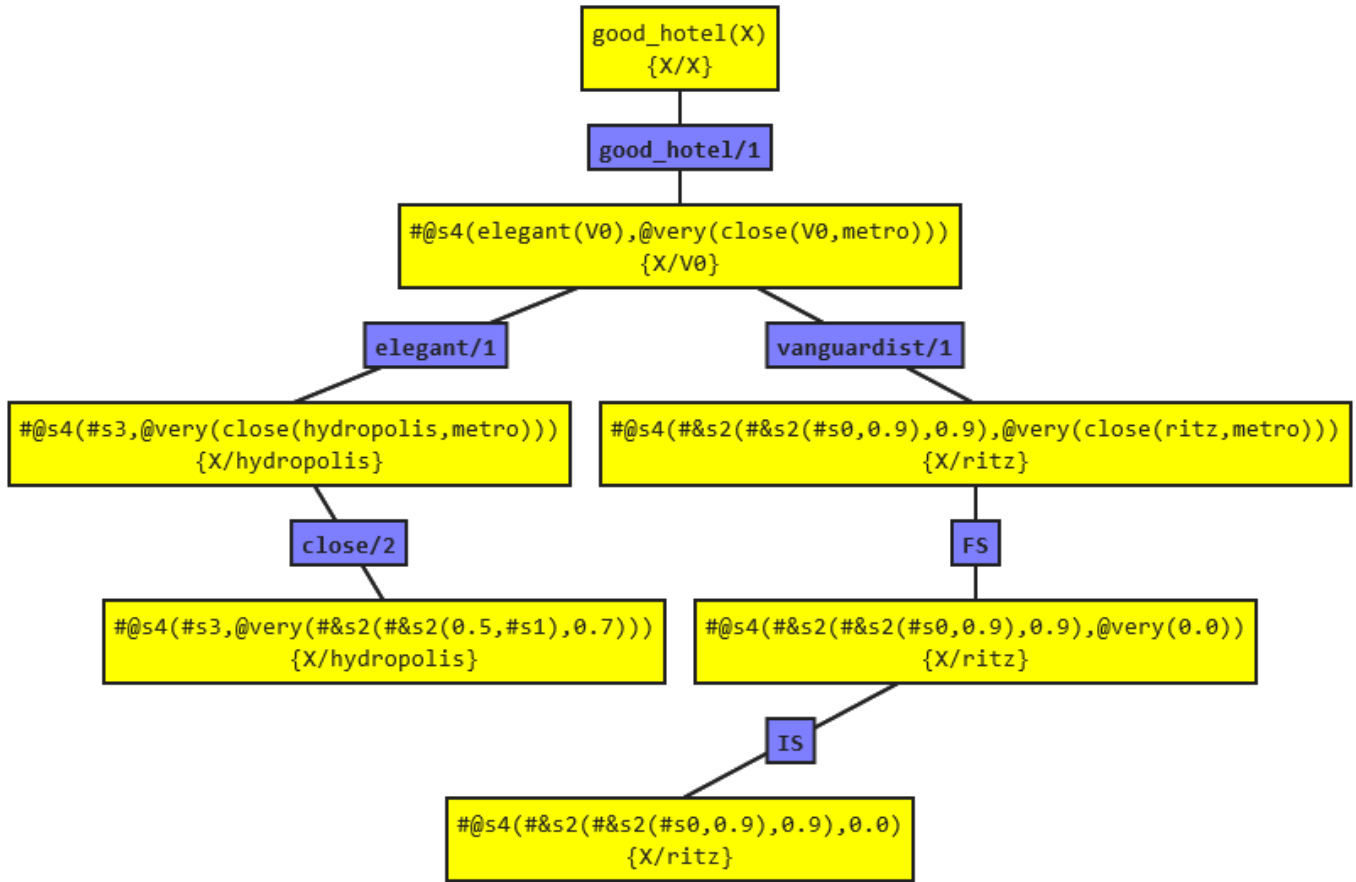


Fig. 3. Screenshot of the FASILL online tool depicting a symbolic derivation tree.

Example 5. Let us build a transformation sequence where each sFASILL program in the sequence is obtained from the immediately preceding one by applying symbolic unfolding, except the initial one $\mathcal{P}_0^\# = \langle \Pi_0^\#, \mathcal{R}^\#, L \rangle$, which, in our case, is the one illustrated in Example 2, that is:

$$\Pi_0^\# = \begin{cases} R_1 : & \text{vanguardist(ritz)} \leftarrow 0.9 \\ R_2 : & \text{elegant(hydropolis)} \leftarrow s_3^\# \\ R_3 : & \text{close(hydropolis, taxi)} \leftarrow 0.7 \\ R_4 : & \text{good_hotel}(x) \leftarrow \\ & \quad \text{@}_{s_4}^\#(\text{elegant}(x), \text{@}_{\text{very}}(\text{close}(x, \text{metro}))) \end{cases}$$

Program $\mathcal{P}_1^\# = \langle \Pi_1^\#, \mathcal{R}^\#, L \rangle$ is obtained after unfolding rule R_4 (with selected atom $\text{elegant}(x)$) by applying a $\overset{SS}{\rightsquigarrow}$ step with rules R_1 and R_2 :

$$\Pi_1^\# = \begin{cases} R_1 : & \text{vanguardist(ritz)} \leftarrow 0.9 \\ R_2 : & \text{elegant(hydropolis)} \leftarrow s_3^\# \\ R_3 : & \text{close(hydropolis, taxi)} \leftarrow 0.7 \\ R_{41} : & \text{good_hotel(ritz)} \leftarrow \text{@}_{s_4}^\#(\\ & \quad \&_{s_2}^\#(\&_{s_2}^\#(s_0^\#, 0.9), 0.9), \\ & \quad \text{@}_{\text{very}}(\text{close(ritz, metro)})) \\ R_{42} : & \text{good_hotel(hydropolis)} \leftarrow \\ & \quad \text{@}_{s_4}^\#(s_3^\#, \text{@}_{\text{very}}(\text{close(hydropolis, metro)})) \end{cases}$$

After unfolding rule R_{41} (with selected atom $\text{close(ritz, metro)}$) by applying a $\overset{FS}{\rightsquigarrow}$ step, we obtain program $\mathcal{P}_2^\# = \langle \Pi_2^\#, \mathcal{R}^\#, L \rangle$:

$$\Pi_2^\# = \begin{cases} R_1 : & \text{vanguardist(ritz)} \leftarrow 0.9 \\ R_2 : & \text{elegant(hydropolis)} \leftarrow s_3^\# \\ R_3 : & \text{close(hydropolis, taxi)} \leftarrow 0.7 \\ R_{41F} : & \text{good_hotel(ritz)} \leftarrow \\ & \quad \text{@}_{s_4}^\#(\&_{s_2}^\#(\&_{s_2}^\#(s_0^\#, 0.9), 0.9), \text{@}_{\text{very}}(0.0)) \\ R_{42} : & \text{good_hotel(hydropolis)} \leftarrow \\ & \quad \text{@}_{s_4}^\#(s_3^\#, \text{@}_{\text{very}}(\text{close(hydropolis, metro)})) \end{cases}$$

When unfolding rule R_{42} (with selected atom $\text{close(hydropolis, metro)}$) by applying a $\overset{SS}{\rightsquigarrow}$ step with rule R_3 , we reach the program $\mathcal{P}_3^\# = \langle \Pi_3^\#, \mathcal{R}^\#, L \rangle$:

$$\Pi_3^\# = \begin{cases} R_1 : & \text{vanguardist(ritz)} \leftarrow 0.9 \\ R_2 : & \text{elegant(hydropolis)} \leftarrow s_3^\# \\ R_3 : & \text{close(hydropolis, taxi)} \leftarrow 0.7 \\ R_{41F} : & \text{good_hotel(ritz)} \leftarrow \\ & \quad \text{@}_{s_4}^\#(\&_{s_2}^\#(\&_{s_2}^\#(s_0^\#, 0.9), 0.9), \text{@}_{\text{very}}(0.0)) \\ R_{423} : & \text{good_hotel(hydropolis)} \leftarrow \\ & \quad \text{@}_{s_4}^\#(s_3^\#, \text{@}_{\text{very}}(\&_{s_2}^\#(\&_{s_2}^\#(0.5, s_1^\#), 0.7))) \end{cases}$$

```

1  vanguardist(ritz) <- 0.9.
2  elegant(hydropolis) <- #s3.
3  close(hydropolis,taxi) <- 0.7.
4  good_hotel(X) <- #@s4(elegant(X), @very(close(X, metro))).
    
```

Linearize program

Extend program

Unfold program

```

1  vanguardist(ritz) <- 0.9.
    
```

```

2  elegant(hydropolis) <- #s3.
    
```

```

3  close(hydropolis,taxi) <- 0.7.
    
```

```

4  good_hotel(X) <- #@s4(elegant(X),@very(close(X,metro))).
    
```

(a) Original sFASILL program before being transformed.

```

1  vanguardist(ritz) <- 0.9.
2  elegant(hydropolis) <- #s3.
3  close(hydropolis,taxi) <- 0.7.
4  good_hotel(hydropolis) <- #@s4(#s3,@very(close(hydropolis,metro))).
5  good_hotel(ritz) <- #@s4(#&s2(#&s2(#s0,0.9),0.9),@very(close(ritz,metro))).
    
```

(b) sFASILL program obtained after unfolding the last program rule.

Fig. 4. The FASILL online tool unfolding a symbolic program.

Finally, by unfolding rule R_{41FI} (with selected expression $@_{very}(0.0)$) after applying a $\overset{IS}{\rightsquigarrow}$ step, we obtain the final program $\mathcal{P}_4^\# = \langle \Pi_4^\#, \mathcal{R}^\#, L \rangle$:

$$\Pi_4^\# = \left\{ \begin{array}{l} R_1 : \quad vanguardist(ritz) \leftarrow 0.9 \\ R_2 : \quad elegant(hydropolis) \leftarrow s_3^\# \\ R_3 : \quad close(hydropolis, taxi) \leftarrow 0.7 \\ R_{41FI} : \quad good_hotel(ritz) \leftarrow \\ \quad \quad \quad @_{s_4}^\#(\&_{s_2}^\#(\&_{s_2}^\#(s_0^\#, 0.9), 0.9), 0.0) \\ R_{423} : \quad good_hotel(hydropolis) \leftarrow \\ \quad \quad \quad @_{s_4}^\#(s_3^\#, @_{very}(\&_{s_2}^\#(\&_{s_2}^\#(0.5, s_1^\#), 0.7))) \end{array} \right.$$

In the previous example, it is easy to see that each program in the sequence produces the same set of sfca's for a given goal but reducing the length of derivations. For instance, the

derivation performed w.r.t. the original program $\mathcal{P}_0^\#$ illustrated in Example 4, can be emulated in the final program $\mathcal{P}_4^\#$ with just one computational step (instead of four) as:

$$\begin{array}{l} \langle good_hotel(x); id \rangle \\ \langle @_{s_4}^\#(\&_{s_2}^\#(\&_{s_2}^\#(s_0^\#, 0.9), 0.9), 0.0); \{x/ritz\} \rangle. \end{array} \overset{SS}{\rightsquigarrow} R_{41FI}$$

IV. TUNING TECHNIQUES FOR sFASILL PROGRAMS

We start this section by summarizing the automated technique for tuning sFASILL programs that we initially presented in [21][22].

Typically, a programmer has a model in mind where some parameters have a clear value. For instance, the truth value of a rule might be statistically determined and, thus, its value can be easily obtained. In other cases, though, the most appropriate values and/or connectives depend on subjective notions and,

thus, programmers do not know how to obtain these values. In a typical scenario, we have an extensive set of *expected* computed answers (i.e., *test cases*), so the programmer can follow a “try and test” strategy. Unfortunately, this is a tedious and time consuming operation. Actually, it might even be impractical when the program should correctly model a large number of test cases.

The first action for initializing the tuning process in the FASILL online tool obviously consists in introducing a set of test cases. The tune area of our online tool is shown in Figure 5a. Each test case appears in a different line with syntax: $r \rightarrow Q$, where r is the desired truth degree for the *fca* associated to query Q (which obviously does not contain symbolic constants). For instance, in our running example we can introduce the following three test cases:

```
0.60 -> modern(hydropolis).
0.45 -> good_hotel(ritz).
0.38 -> good_hotel(hydropolis).
```

Then, users need to select a tuning method and click on the Tune program button to proceed with the tuning process. The precision of the technique depends on the set of symbolic substitutions considered at tuning time. So, for assigning values to the symbolic constants, our tool takes into account all the truth values defined on a *members/1* predicate (which in our case is declared as *members*([0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1])) as well as the set of connectives defined in the lattice associated to the program, which in our running example coincides with the three conjunction and disjunction connectives based on the so-called *Product*, *Gödel* and *Lukasiewicz* logics, as shown in Figure 1, and the arithmetic/geometric average aggregators defined in Example 1. Obviously, the larger the domain of values and connectives is, the more precise the results are (but the algorithm is more expensive, of course).

As we have seen in this work, symbolic programs can be seen as *partial specifications* of fuzzy programs, where some elements on program rules have not been clearly identified yet. Here, we assume that test cases are provided a priori by users in order to establish well known information about a particular domain and, for this reason, they will usually be expressed as ground atoms with associated (desired) truth degrees. It is easy to think that the three test cases used in our running example (see Figure 5a) refers to user’s preferences that could be collected, for instance, by means of satisfaction questionnaires or any other alternative kind of statistically-based analysis. However, our system admits test cases non necessarily based on ground atoms,

For tuning an *sFASILL* program, we have implemented several methods in previous works, including some very efficient versions based on SAT/SMT solvers [38], which are out of the scope of this paper since they only work with connectives whose truth functions are linear. In this work, we focus on the two more general methods described in [21], which exhibit different run-times depending on when and where symbolic substitutions are applied (to symbolic programs or to *sfca*’s):

- **Basic:** The basic method is based on applying each symbolic substitution to the original *sFASILL* program and then fully executing the resulting instantiated FASILL programs. This method is represented by Algorithm 1.
- **Symbolic:** In this version, symbolic substitutions are directly applied to *sfca*’s and thus, only interpretive (but neither successful nor failure) steps are repeatedly executed on the instantiated *fca*’s. This method is represented by Algorithm 2.

Both algorithms use thresholding techniques for prematurely disregarding computations leading to non significant solutions. In [19], [39], [40], [41] we document some interesting results achieved in our research group when designing sophisticated tools for manipulating fuzzy logic programs and, now, we apply the same techniques for improving the efficiency of the basic and symbolic tuning methods under comparison in this section.

Algorithm 1: Basic Tuning for FASILL programs.

Data: A symbolic program $\mathcal{P}^\#$ and a set of test cases

$\{\langle \mathcal{G}_1, v_1 \rangle, \dots, \langle \mathcal{G}_k, v_k \rangle\}$.

Result: A symbolic substitution Θ .

Consider a finite number of symbolic substitutions

$\Theta_1, \dots, \Theta_n$ for $\text{sym}(\mathcal{P}^\#) \cup \bigcup_{i=1}^k \text{sym}(\mathcal{G}_i)$;

$\tau \leftarrow +\infty$;

foreach $i \in \{1, \dots, n\}$ **do**

$\mathcal{P}_i \leftarrow \mathcal{P}^\# \Theta_i$;

$\epsilon \leftarrow 0$;

foreach $j \in \{1, \dots, k\}$ **do**

 Compute the *fca* $\langle \mathcal{G}_j \Theta_i; id \rangle \rightsquigarrow^* \langle v_{i,j}; \theta_{i,j} \rangle$ in

\mathcal{P}_i ;

$\epsilon \leftarrow \epsilon + d(v_{i,j}, v_j)$;

if $\epsilon \geq \tau$ **then**

break;

end

end

if $\epsilon < \tau$ **then**

$\tau \leftarrow \epsilon$;

$\Theta \leftarrow \Theta_i$;

end

end

return Θ ;

In the symbolic algorithm and, as we detected in [21], we must take care when attaching concrete values to the symbolic constants appearing in the symbolic similarity relations. Beyond the simpler case of assigning the \perp truth degree to at least a symbolic constant, it is also possible to conceive other non safe symbolic substitutions linking symbolic constant to values bigger than \perp . For instance, this is the case of $\Theta = \{s_1^\# / 0.4, \&_{s_2^\#} / \&_{1uka}, \dots\}$ in our running example, because if, in particular, we apply this symbolic substitution to $\mathcal{R}^\#(taxi, metro) = 0.5 \&_{s_2^\#} s_1^\#$ we have that $\llbracket \mathcal{R}^\#(taxi, metro) \Theta \rrbracket = 0.5 \&_{1uka} 0.4 = \max(0, 0.5 + 0.4 - 1) = \max(0, -0.1) = 0$, which implies that Θ is not a

Algorithm 2: Symbolic Tuning for FASILL programs.

Data: A symbolic program $\mathcal{P}^\#$ and a set of test cases $\{\langle \mathcal{G}_1, v_1 \rangle, \dots, \langle \mathcal{G}_k, v_k \rangle\}$.

Result: A symbolic substitution Θ .

foreach $i \in \{1, \dots, k\}$ **do**
 | Compute the sfca $\langle \mathcal{G}_i, id \rangle \rightsquigarrow^* \langle \mathcal{G}'_i, \theta_i \rangle$ in $\mathcal{P}^\#$;
end

Consider a finite number of symbolic substitutions $\Theta_1, \dots, \Theta_n$ for $\bigcup_{i=1}^k \text{sym}(\mathcal{G}'_i)$;
 $\tau \leftarrow +\infty$;

foreach $i \in \{1, \dots, n\}$ **do**
 | **if** Θ_i is \mathcal{R} -safe **then**
 | | $\epsilon \leftarrow 0$;
 | | **foreach** $j \in \{1, \dots, k\}$ **do**
 | | | Compute the fca $\langle \mathcal{G}'_j \Theta_i; \theta_j \rangle \stackrel{\text{IS}^*}{\rightsquigarrow} \langle v_{i,j}; \theta_j \rangle$
 | | | in $\mathcal{P}^\#$;
 | | | $\epsilon \leftarrow \epsilon + d(v_{i,j}, v_j)$;
 | | | **if** $\epsilon \geq \tau$ **then**
 | | | | **break**;
 | | | **end**
 | | | **end**
 | | | **if** $\epsilon < \tau$ **then**
 | | | | $\tau \leftarrow \epsilon$;
 | | | | $\Theta \leftarrow \Theta_i$;
 | | | **end**
 | | **end**
end

end
return Θ ;

safe symbolic substitution w.r.t. $\mathcal{R}^\#$. This example justifies why, in Algorithm 2, we use the concept of *safe symbolic substitution* introduced in [21] (Definition 7). In essence, given a symbolic similarity relation $\mathcal{R}^\#$ on a domain \mathcal{U} and a symbolic substitution Θ , we say that Θ is a *safe symbolic substitution* w.r.t. $\mathcal{R}^\#$ if, for all $x, y \in \mathcal{U}$ such that $\mathcal{R}^\#(x, y)$ is an $L^\#$ -expression containing at least a symbolic constant, then $\llbracket \mathcal{R}^\#(x, y) \Theta \rrbracket \neq \perp$.

As seen in Figure 5b, the system reports the best symbolic substitution obtained after performing the tuning process, as well as its associated deviation. In our case, the best symbolic substitution is $\Theta = \{s_0^\#/0.9, s_1^\#/0.4, \&_{s_2}^\#/\&_{\text{goodel}}, s_3^\#/0.6, @_{s_4}^\#/@_{\text{aver}}\}$, which has no deviation w.r.t. the three test cases (although this is not always the general case) and hence, once Θ is applied to the original sFASILL program and, after executing a goal like `good_hotel(X)` w.r.t. the final, tuned FASILL program, we obtain the fca's $\langle 0.45, \{X/\text{ritz}\} \rangle$ and $\langle 0.38, \{X/\text{hydropolis}\} \rangle$, while the execution of goal `modern(hydropolis)` returns $\langle 0.60, \{\} \rangle$, which coincide with the preferences expressed in the three test cases, as wanted.

Let us finish this section by mentioning that the solutions achieved by both the basic and symbolic algorithms is always

the same for all programs described in the previous section (it doesn't matter if they have been more or less unfolded), but the time required to reach the best symbolic substitutions can drastically change, as we are going to compare and explain in the following section.

V. COMBINING UNFOLDING AND TUNING TECHNIQUES. PERFORMANCE, EVALUATION AND DISCUSSION

Observe in Figure 5b that the FASILL online tool also reports the time required to find the solution of a tuning program, which in our running example is close to 3.7 seconds when applying the symbolic tuning algorithm, but this time almost doubles when selecting the basic method. Let us start this section by explaining the tuning session sketched in Table I, where each row refers to a different symbolic substitution fully checked at tuning time and the first five columns indicate the concrete values and connectives linked to the five symbolic constants in our running example. Columns labelled with t_i and ϵ_i , $1 \leq i \leq 3$, indicate the achieved truth degree and partial deviation, respectively, associated to the evaluation of each one of the three test cases, being the last column ϵ the global deviation produced by the corresponding symbolic substitution.

Since there are two possible aggregators to select for $@_{s_4}^\#$, three conjunctions for mapping $\&_{s_2}^\#$ and eleven numeric values as choices for $s_0^\#, s_1^\#$ and $s_3^\#$, the system considers 7986 symbolic substitution as candidates. Obviously, we can not display so many rows in the table but, fortunately, there is no need to do so thanks to the benefits introduced by thresholding when evaluating the test cases. In fact, the small subset of rows in the table is sufficient to illustrate our technique, since only these symbolic substitutions are the only ones that both tuning algorithms (basic and symbolic) apply to all test cases for fully evaluating them (any other candidate is abruptly ruled out as soon as possible). Technically, observe in both tuning algorithms that threshold τ (whose initial value is $+\infty$), is dynamically updated in sentence “**if** ($\epsilon < \tau$) ...” whenever the systems reaches a symbolic substitution Θ_i whose global deviation is better than the one of the previously proposed as solution Θ , and also τ is used in sentence “**if** ($\epsilon \geq \tau$) ...” for prematurely disregarding a concrete symbolic substitution without evaluating all test cases whenever the deviation accumulated by some of them is excessive. So, since the table only displays symbolic substitutions improving the deviation achieved by previous candidates, observe, for instance, that all intermediate candidates between Θ_{923} and Θ_{1043} are omitted because their accumulated deviations (without the mandatory need of evaluating all test cases) are bigger than the one of Θ_{923} , and this last one is also bigger than the deviation associated to Θ_{1043} .

Until now, we have considered unfolding and tuning operations as independent techniques with clearly different objectives: the first one to improve the efficiency of programs, and the second one to calibrate their rules. However, the unfolding transformation can also be used to improve the tuning process,

✎ Test cases

```
1 0.60 -> modern(hydropolis).
2 0.45 -> good_hotel(ritz).
3 0.38 -> good_hotel(hydropolis).
```

FASILL – Symbolic method ▾

Tune program

FASILL – Symbolic method

FASILL – Basic method

SMT – Boolean lattice

SMT – Real lattice

SMT – Unit lattice

(a) Screenshot of the online tool for starting a tuning process.

🔍 Symbolic substitution

```
1 best symbolic substitution: {#s4/@aver,#s2/&godel,#s0/0.9,#s1/0.4,#s3/0.6}
2 deviation: 0.0
3 execution time: 3699 milliseconds
```

(b) Screenshot of the online tool after ending a tuning process.

Fig. 5. The FASILL online tool tuning a symbolic program.

in order to approximate the run-time of the basic method to the symbolic one.

Coming back again to our tuning example, when tuning both the original and unfolded programs of examples 2 and 5, we obviously obtain the same symbolic substitution, apart from the fact that the unfolded program will always run faster than the original one. Moreover, the tuning-time of the basic method is considerably reduced when applied to the transformed program. This is due to the fact that, instead of using several successful/failure step w.r.t. the original program, after unfolding it, the basic method only needs to apply just one $\overset{SS/FS}{\rightsquigarrow}$ step to get the fca of each test case after applying each symbolic substitution to the entire program.

Anyway, the basic method is still slightly slower than the symbolic one, even with the unfolded version of the program. That is because the basic method has to perform a $\overset{SS/FS}{\rightsquigarrow}$ step for each symbolic substitution, while the symbolic method calculates the sfca's (also in one $\overset{SS/FS}{\rightsquigarrow}$ step) only once and stores them. So, if we consider n symbolic substitutions and k test cases for the tuning process, the basic method (with the unfolded program) should compute $(n - 1) * k$ more admissible steps than the symbolic method. All in all, both tuning algorithms preceded by the unfolding pre-process require

approximately the same time than the symbolic algorithm in isolation (the basic alone doubles such time, as commented before) in our running example. But in order to confirm this property in a more complex situation, we have prepared the following benchmarks.

Tables II and III summarize the results of an experimental evaluation (each cell refers to the average of 10 executions using a desktop computer equipped with 4,00 GB RAM and i3-2310M CPU @ 2.10 GHz.) of the tuning techniques preceded by several unfolding iterations. Here, the same sFASILL program is tuned with both basic and symbolic algorithms, varying the numbers of unfolding operations and explored symbolic substitutions. To do this, we consider a general rule of the form $p \leftarrow q \wedge \dots \wedge q \wedge s^\#$ (containing 100 instances of atom q and only a symbolic constant $s^\#$), along with a fact $q \leftarrow \top$; and we introduce a single test case: $\top \rightarrow p$. Furthermore, in all executions, the only symbolic substitution that produces a deviation of 0 is considered the last one in the search space to ensure that both methods explore exactly the maximum number of symbolic substitutions.

Focusing on the first row in both tables, referred to the tuning time associated to the manipulation of the original program when considering from 10 to 1000 different symbolic

TABLE I. Search for the best symbolic substitution by the FASILL system, where t_i and ϵ_i denote the truth degree and deviation for the i -th test case. Only symbolic substitutions that improve the error made by previous substitutions in the search process are shown.

Θ	@# _{s4}	&# _{s2}	s# ₀	s# ₁	s# ₃	t_1	ϵ_1	t_2	ϵ_2	t_3	ϵ_3	ϵ_{total}
Θ_{309}	@aver	&luka	0.2	0.6	0.0	0.0	0.6	0.0	0.45	0.0	0.38	1.4300
Θ_{310}	@aver	&luka	0.2	0.6	0.1	0.0	0.6	0.0	0.45	0.05	0.33	1.3800
Θ_{311}	@aver	&luka	0.2	0.6	0.2	0.0	0.6	0.0	0.45	0.1	0.28	1.3300
Θ_{312}	@aver	&luka	0.2	0.6	0.3	0.0	0.6	0.0	0.45	0.15	0.23	1.2800
Θ_{313}	@aver	&luka	0.2	0.6	0.4	0.0	0.6	0.0	0.45	0.2	0.18	1.2300
Θ_{314}	@aver	&luka	0.2	0.6	0.5	0.0	0.6	0.0	0.45	0.25	0.13	1.1800
Θ_{315}	@aver	&luka	0.2	0.6	0.6	0.0	0.6	0.0	0.45	0.3	0.08	1.1300
Θ_{316}	@aver	&luka	0.2	0.6	0.7	0.0	0.6	0.0	0.45	0.35	0.03	1.0800
Θ_{317}	@aver	&luka	0.2	0.6	0.8	0.0	0.6	0.0	0.45	0.4	0.02	1.0700
Θ_{318}	@aver	&luka	0.2	0.6	0.9	0.1	0.5	0.0	0.45	0.45	0.07	1.0200
Θ_{319}	@aver	&luka	0.2	0.6	1.0	0.2	0.4	0.0	0.45	0.5	0.12	0.9700
Θ_{438}	@aver	&luka	0.3	0.6	0.8	0.1	0.5	0.05	0.4	0.4	0.02	0.9200
Θ_{439}	@aver	&luka	0.3	0.6	0.9	0.2	0.4	0.05	0.4	0.45	0.07	0.8700
Θ_{440}	@aver	&luka	0.3	0.6	1.0	0.3	0.3	0.05	0.4	0.5	0.12	0.8200
Θ_{559}	@aver	&luka	0.4	0.6	0.8	0.2	0.4	0.1	0.35	0.4	0.02	0.7700
Θ_{560}	@aver	&luka	0.4	0.6	0.9	0.3	0.3	0.1	0.35	0.45	0.07	0.7200
Θ_{561}	@aver	&luka	0.4	0.6	1.0	0.4	0.2	0.1	0.35	0.5	0.12	0.6700
Θ_{680}	@aver	&luka	0.5	0.6	0.8	0.3	0.3	0.15	0.3	0.4	0.02	0.6200
Θ_{681}	@aver	&luka	0.5	0.6	0.9	0.4	0.2	0.15	0.3	0.45	0.07	0.5700
Θ_{682}	@aver	&luka	0.5	0.6	1.0	0.5	0.1	0.15	0.3	0.5	0.12	0.5200
Θ_{801}	@aver	&luka	0.6	0.6	0.8	0.4	0.2	0.2	0.25	0.4	0.02	0.4700
Θ_{802}	@aver	&luka	0.6	0.6	0.9	0.5	0.1	0.2	0.25	0.45	0.07	0.4200
Θ_{803}	@aver	&luka	0.6	0.6	1.0	0.6	0.0	0.2	0.25	0.5	0.12	0.3700
Θ_{922}	@aver	&luka	0.7	0.6	0.8	0.5	0.1	0.25	0.2	0.4	0.02	0.3200
Θ_{923}	@aver	&luka	0.7	0.6	0.9	0.6	0.0	0.25	0.2	0.45	0.07	0.2700
Θ_{1043}	@aver	&luka	0.8	0.6	0.8	0.6	0.0	0.3	0.15	0.4	0.02	0.1700
Θ_{1163}	@aver	&luka	0.9	0.6	0.7	0.6	0.0	0.35	0.1	0.35	0.03	0.1300
Θ_{1196}	@aver	&luka	0.9	0.9	0.7	0.6	0.0	0.35	0.1	0.355	0.025	0.1250
Θ_{1207}	@aver	&luka	0.9	1.0	0.7	0.6	0.0	0.35	0.1	0.37	0.01	0.1100
Θ_{2603}	@aver	&prod	1.0	0.5	0.6	0.6	0.0	0.405	0.045	0.3153	0.0647	0.1097
Θ_{2614}	@aver	&prod	1.0	0.6	0.6	0.6	0.0	0.405	0.045	0.322	0.058	0.1029
Θ_{2625}	@aver	&prod	1.0	0.7	0.6	0.6	0.0	0.405	0.045	0.33	0.05	0.0950
Θ_{2636}	@aver	&prod	1.0	0.8	0.6	0.6	0.0	0.405	0.045	0.3392	0.0408	0.0858
Θ_{2647}	@aver	&prod	1.0	0.9	0.6	0.6	0.0	0.405	0.045	0.3496	0.0304	0.0754
Θ_{2658}	@aver	&prod	1.0	1.0	0.6	0.6	0.0	0.405	0.045	0.3612	0.0188	0.0638
Θ_{3681}	@aver	&godel	0.8	0.4	0.6	0.6	0.0	0.4	0.05	0.38	0.0	0.0500
Θ_{3791}	@aver	&godel	0.9	0.3	0.6	0.6	0.0	0.45	0.0	0.345	0.035	0.0350
Θ_{3802}	@aver	&godel	0.9	0.4	0.6	0.6	0.0	0.45	0.0	0.38	0.0	0.0000

TABLE II. Average runtime (in milliseconds) of the basic tuning method after 10 executions, based on the number of unfolding steps (k) and the number of considered symbolic substitutions.

k	Time (ms)					
	10 Θ	50 Θ	100 Θ	250 Θ	500 Θ	1000 Θ
0	168	781	1359	3515	6946	13772
1	159	693	1372	3453	6768	13578
5	128	672	1356	3681	6593	13059
10	140	634	1265	3334	6368	13006
25	125	562	1140	2825	5640	11425
50	90	428	853	2115	4515	8456
75	75	340	518	1406	2531	5028
100	15	53	106	278	540	1065

TABLE III. Average runtime (in milliseconds) of symbolic tuning method after 10 executions, based on the number of unfolding steps (k) and the number of considered symbolic substitutions.

k	Time (ms)					
	10 Θ	50 Θ	100 Θ	250 Θ	500 Θ	1000 Θ
0	56	65	115	268	578	1240
1	50	65	115	268	540	1043
5	22	59	115	268	518	1037
10	25	62	97	256	534	1043
25	21	62	115	309	525	1075
50	18	59	112	265	522	1159
75	28	56	109	253	512	1034
100	15	53	106	243	515	1025

substitutions, the basic algorithm ranges from 168 to 13772 milliseconds in Table II, which largely contrasts with the rank between 56 and 1240 exhibited by the symbolic method in Table III. Also, the last column in the basic case presents drastic time reductions when comparing the tuning time of the original program with respect to its different unfolded versions, being this measure dramatically reduced (approximately) thirteen times when tuning the last transformed program obtained after unfolding one hundred times the initial one. In contrast to this, note that the tuning time of the symbolic algorithm is not especially affected by unfolding transformations, varying

from 1240 to 1025 milliseconds in the last column of Table III and being this last value almost the same (1065) than the one in the cell at the right-bottom corner of Table II. This effect is reinforced in Figure 6, which evidences in a more graphical way that the application of unfolding before tuning a program is not relevant for the symbolic strategy, but the efficiency of the basic method notably increases, which confirms our expectations.

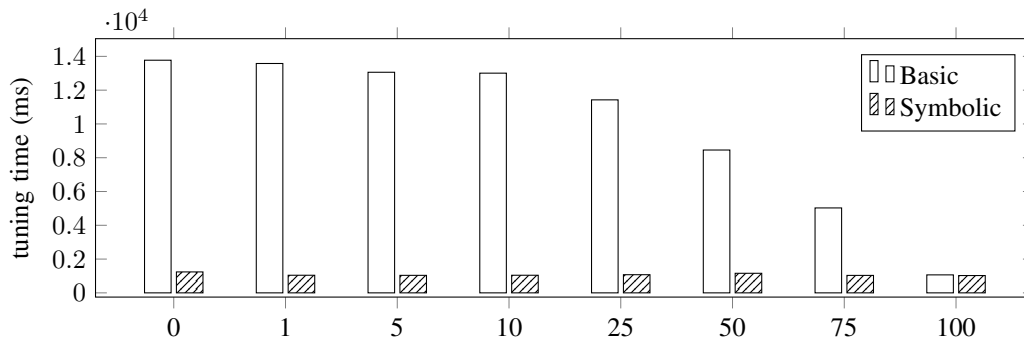


Fig. 6. Comparison of tuning algorithms based on the number of unfolding steps performed over the program.

VI. CONCLUSION AND FUTURE WORK

Coping with symbolic similarity relations, the symbolic extension of the FASILL language we introduced in [22], is the basis of the effective unfolding technique for sFASILL programs we have very recently presented in [1]. The new transformation surpasses both the similarity-based (but non-symbolic) unfolding of [30][31], as well as the symbolic (but not dealing yet with similarities) operation of [32][33], thus permitting the optimization of sFASILL programs in an unified, similarity-based symbolic framework.

The present work also considers tuning techniques for extending [1], having in mind that sFASILL programs can be seen as templates to be fulfilled with concrete fuzzy values/operators for producing tuned FASILL programs which satisfy users wishes. In this paper, we have firstly collected from [38][21] two semi-automatic tuning engines (basic and symbolic) for customizing symbolic programs and then, we have combined them with the symbolic unfolding technique of [1]. We have implemented these techniques in a freely available tool [34], which has been used for developing interesting benchmarks and analyzing the good performance of the mixed techniques. Our experiments reveal that, when manipulating sFASILL programs, the sequence “unfolding plus tuning” is preferable than the one of “tuning plus unfolding”. To summarize, the benefits of applying symbolic unfolding on a sFASILL program before tuning it, are: 1) the basic method applied on unfolded programs exhibits an efficiency close to the symbolic algorithm (with or without an unfolding pre-process) and 2) both the execution and tuning times are significantly improved for the resulting, unfolded plus tuned, FASILL program.

Some pending related tasks for the near future consist in exploring the synergies between our approach and machine learning strategies (including neural networks, as we managed in [27]), fuzzy SMT/SAT solvers (see our previous experiences in [38]), etc. As ongoing work, we are nowadays developing the formal proofs that reinforce the correctness of symbolic unfolding under certain safe applicability conditions.

ACKNOWLEDGMENT

This work has been partially supported by the EU (FEDER), the State Research Agency (AEI) of the Spanish Ministry of

Science and Innovation under grant PID2019-104735RB-C42 (SAFER).

REFERENCES

- [1] G. Moreno and J. A. Riaza, “Symbolic unfolding of similarity-based fuzzy logic programs,” in *The 2024 IARIA Annual Congress on Frontiers in Science, Technology, Services, and Applications*. IARIA Congress 2024, ThinkMind Digital Library, 2019, pp. 121–125. [Online]. Available: https://personales.upv.es/thinkmind/dl/conferences/iariacongress/iaria_congress_2024/iaria_congress_2024_2_160_50079.pdf
- [2] L. A. Zadeh, “Fuzzy sets,” *Information and Control*, vol. 8, pp. 338–353, 1965. [Online]. Available: <http://www-bisc.cs.berkeley.edu/Zadeh-1965.pdf>
- [3] J. W. Lloyd, *Foundations of Logic Programming*. Springer, 1987.
- [4] P. Vojtáš, “Fuzzy Logic Programming,” *Fuzzy Sets and Systems*, vol. 124, no. 1, pp. 361–370, 2001.
- [5] F. Formato, G. Gerla, and M. I. Sessa, “Similarity-based unification,” *Fundamenta Informaticae*, vol. 41, no. 4, pp. 393–414, 2000.
- [6] M. I. Sessa, “Approximate reasoning by similarity-based SLD resolution,” *Theoretical Computer Science*, vol. 275, no. 1-2, pp. 389–426, 2002.
- [7] F. Arcelli, “Likelog for flexible query answering,” *Soft Computing*, vol. 7, no. 2, pp. 107–114, 2002.
- [8] P. Julián-Iranzo and C. Rubio, “A declarative semantics for bousi~prolog,” in *Proc. of 11th Int. ACM SIGPLAN Conference on Principles and Practice of Declarative Programming, PPDP’09, Coimbra, Portugal*. ACM, 2009, pp. 149–160.
- [9] C. Rubio-Manzano and P. Julián-Iranzo, “A fuzzy linguistic prolog and its applications,” *Journal of Intelligent and Fuzzy Systems*, vol. 26, no. 3, pp. 1503–1516, 2014.
- [10] R. Lee, “Fuzzy Logic and the Resolution Principle,” *Journal of the ACM*, vol. 19, no. 1, pp. 119–129, 1972.
- [11] M. Ishizuka and N. Kanai, “Prolog-ELF Incorporating Fuzzy Logic,” in *Proceedings of the 9th Int. Joint Conference on Artificial Intelligence, IJCAI’85*, A. K. Joshi, Ed. Morgan Kaufmann, 1985, pp. 701–703.
- [12] D. Li and D. Liu, *A fuzzy Prolog database system*. John Wiley & Sons, Inc., 1990.
- [13] M. Kifer and V. S. Subrahmanian, “Theory of generalized annotated logic programming and its applications,” *Journal of Logic Programming*, vol. 12, pp. 335–367, 1992.
- [14] J. F. Baldwin, T. P. Martin, and B. W. Pilsworth, *FriL- Fuzzy and Evidential Reasoning in Artificial Intelligence*. John Wiley & Sons, Inc., 1995.
- [15] M. Rodríguez-Artalejo and C. Romero-Díaz, “Quantitative logic programming revisited,” in *Proc. of 9th Functional and Logic Programming Symposium, FLOPS’08*, J. Garrigue and M. Hermenegildo, Eds. LNCS, 4989, Springer Verlag, 2008, pp. 272–288.
- [16] S. Guadarrama, S. Muñoz, and C. Vaucheret, “Fuzzy Prolog: A new approach using soft constraints propagation,” *Fuzzy Sets and Systems*, vol. 144, no. 1, pp. 127–150, 2004.
- [17] S. Muñoz, V. P. Ceruelo, and H. Strass, “Rfuzzy: Syntax, semantics and implementation details of a simple and expressive fuzzy tool over prolog,” *Information Sciences*, vol. 181, no. 10, pp. 1951–1970, 2011.

- [18] J. Medina, M. Ojeda-Aciego, and P. Vojtáš, "Similarity-based Unification: a multi-adjoint approach," *Fuzzy Sets and Systems*, vol. 146, pp. 43–62, 2004.
- [19] P. Julián, G. Moreno, and J. Penabad, "Thresholded semantic framework for a fully integrated fuzzy logic language," *J. Log. Algebr. Meth. Program.*, vol. 93, pp. 42–67, 2017. [Online]. Available: <https://doi.org/10.1016/j.jlmp.2017.08.002>
- [20] G. Moreno, J. Penabad, J. A. Riaza, and G. Vidal, "Symbolic execution and thresholding for efficiently tuning fuzzy logic programs," in *Logic-Based Program Synthesis and Transformation, Proc. of the 26th International Symposium LOPSTR 2016*. LNCS, 10184, Springer, 2016, pp. 131–147. [Online]. Available: https://doi.org/10.1007/978-3-319-63139-4_8
- [21] G. Moreno and J. A. Riaza, "A safe and effective tuning technique for similarity-based fuzzy logic programs," in *Advances in Computational Intelligence - 16th Int. Work-Conference on Artificial Neural Networks, IWANN 2021*, vol. LNCS 12861. Springer, 2021, pp. 190–201. [Online]. Available: https://doi.org/10.1007/978-3-030-85030-2_16
- [22] G. Moreno and J. A. Riaza, "Symbolic similarity relations for tuning fully integrated fuzzy logic programs," in *Rules and Reasoning - Proc. of 4th Int. Joint Conference, RuleML+RR 2020*, vol. LNCS 12173. Springer, 2020, pp. 150–158. [Online]. Available: https://doi.org/10.1007/978-3-030-57977-7_11
- [23] L. D. Raedt and A. Kimmig, "Probabilistic (logic) programming concepts," *Mach. Learn.*, vol. 100, no. 1, pp. 5–47, 2015. [Online]. Available: <https://doi.org/10.1007/s10994-015-5494-z>
- [24] F. Riguzzi and T. Swift, "The PITA system: Tabling and answer subsumption for reasoning under uncertainty," *Theory Pract. Log. Program.*, vol. 11, no. 4-5, pp. 433–449, 2011. [Online]. Available: <https://doi.org/10.1017/S147106841100010X>
- [25] K. F. Sagonas, T. Swift, and D. S. Warren, "XSB as an Efficient Deductive Database Engine," in *Proc. of ACM SIGMOD International Conference on Management of Data*. ACM Press, 1994, pp. 442–453.
- [26] J. Almendros-Jiménez, A. Becerra-Terón, G. Moreno, and J. A. Riaza, "Tuning fuzzy sparql queries," *International Journal of Approximate Reasoning*, vol. 170, p. 109209, 2024. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85191942148&doi=10.1016%2fj.ijar.2024.109209&partnerID=40&md5=42d6833e308f5a0eba0e6240657ae987>
- [27] G. Moreno, J. Pérez, and J. A. Riaza, "Fuzzy logic programming for tuning neural networks," in *Rules and Reasoning - Proc. of Third International Joint Conference, RuleML+RR 2019*, vol. LNCS 11784. Springer, 2019, pp. 190–197.
- [28] A. Pettorossi and M. Proietti, "Rules and Strategies for Transforming Functional and Logic Programs," *ACM Computing Surveys*, vol. 28, no. 2, pp. 360–414, 1996.
- [29] H. Tamaki and T. Sato, "Unfold/Fold Transformations of Logic Programs," in *Proc. of Second Int'l Conf. on Logic Programming*, S. Tärnlund, Ed., 1984, pp. 127–139.
- [30] P. Julián-Iranzo, G. Moreno, and J. A. Riaza, "Seeking a safe and efficient similarity-based unfolding rule," *Int. J. Approx. Reason.*, vol. 163, p. 109038, 2023. [Online]. Available: <https://doi.org/10.1016/j.ijar.2023.109038>
- [31] P. Julián-Iranzo, G. Moreno, and J. A. Riaza, "Some properties of substitutions in the framework of similarity relations," *Fuzzy Sets Syst.*, vol. 465, p. 108510, 2023. [Online]. Available: <https://doi.org/10.1016/j.fss.2023.03.013>
- [32] G. Moreno, J. Penabad, and J. A. Riaza, "Symbolic unfolding of multi-adjoint logic programs," in *Trends in Mathematics and Computational Intelligence*. Studies in Computational Intelligence, Springer International Publishing, (extended version of a previous paper presented at ESCIM'17), 2019, pp. 43–51. [Online]. Available: https://doi.org/10.1007/978-3-030-00485-9_5
- [33] G. Moreno and J. A. Riaza, "An online tool for unfolding symbolic fuzzy logic programs," in *Advances in Computational Intelligence - Proc. of 15th International Work-Conference on Artificial Neural Networks (Part II), IWANN 2019*, vol. LNCS 11507. Springer, 2019, pp. 475–487. [Online]. Available: https://doi.org/10.1007/978-3-030-20518-8_40
- [34] G. Moreno and J. A. Riaza, "Fasill: Sandbox," in <https://dectau.uclm.es/fasill/sandbox>, Accessed: 2024-11-30.
- [35] P. Julián-Iranzo, "A procedure for the construction of a similarity relation," in *Proc. of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU 2008)*, June 22-27, Torremolinos (Málaga), Spain. U. Málaga (ISBN 978-84-612-3061-7), 2008, pp. 489–496.
- [36] A. Kandel and L. Yellowitz, "Fuzzy chains," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. SMC-4, no. 5, pp. 472–475, 1974.
- [37] H. Naessens, H. D. Meyer, and B. D. Baets, "Algorithms for the computation of t-transitive closures," *IEEE Trans. Fuzzy Systems*, vol. 10, no. 4, pp. 541–551, 2002.
- [38] G. Moreno and J. A. Riaza, "Using SAT/SMT Solvers for Efficiently Tuning Fuzzy Logic Programs," in *2020 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2020, Glasgow, UK*. IEEE, 2020, pp. 1–8.
- [39] P. Julián-Iranzo, J. Medina, G. Moreno, and M. Ojeda, "Thresholded tabulation in a fuzzy logic setting," *ENTCS*, vol. 248, pp. 115–130, 2009.
- [40] P. Julián-Iranzo, J. Medina, G. Moreno, and M. Ojeda, "Efficient thresholded tabulation for fuzzy query answering," *Studies in Fuzziness and Soft Computing (Foundations of Reasoning under Uncertainty)*, vol. 249, pp. 125–141, 2010.
- [41] P. Julián-Iranzo, G. Moreno, and J. Penabad, "Efficient reductants calculi using partial evaluation techniques with thresholding," *Electronic Notes in Theoretical Computer Science, Elsevier Science*, vol. 188, pp. 77–90, 2007.

Evaluation of AI Learning Materials Using Physical Computing

Toshiyasu Kato

Department of Information and Media Engineering
Nippon Institute of Technology
Minamisaitama-gun - Saitama, Japan
email: katoto@nit.ac.jp

Yuto Chino

Department of Technology
Fuchu City Fuchu 6th Junior High School
Fuchu - Tokyo, Japan
email: edu.yutochino@gmail.com

Abstract—AI services, including generative AI, have become widespread globally. We are using Artificial Intelligence daily. However, without proper knowledge, users may not achieve the desired results, and there is a risk of inaccuracy. Educational institutions are beginning to establish the groundwork for AI learning. Due to broad learning standards, however, there is few educational materials that cover the fundamental knowledge and skills of using AI. To address this problem, we have developed educational materials that enable basic learning about the mechanisms of AI and motivate learning. For this purpose, we utilize physical computing. This paper reports on the process from the composition of learning standards to the development of educational materials. Furthermore, an experiment to evaluate the effectiveness was conducted.

Keywords—AI learning materials; physical computing; learning standards.

I. INTRODUCTION

This paper is an extension of work originally presented in The First International Conference on IoT-AI 2024 (IARIA) [1].

The proliferation of Artificial Intelligence (AI) services, including generative AI, has made artificial intelligence a familiar presence worldwide. A study conducted by a research group of Massachusetts Institute of Technology involved a task where participants used ChatGPT, one of the generative AIs, to write texts specialized in their areas of expertise. The results showed that the group using ChatGPT reduced the average time required by 40% and increased the quality of output by 18% [2]. In Japan, the Cabinet Office has committed to educational reforms by defining "Mathematics, Data Science, and AI" as the new basics of reading, writing, and arithmetic for the digital society in its AI Strategy 2019 [3]. Acquiring AI literacy is becoming indispensable for thriving in the digital society.

However, there are few examples of educational materials that enable the learning of foundational AI knowledge and skills. We can observe only a few classroom practices in middle school technology and high school industrial arts courses [4] [5]. Consequently, there is a lack of materials that facilitate active learning by students. Furthermore, although practical lessons for acquiring AI literacy are being conducted in primary and secondary education, there is an insufficient learning foundation for instructors.

Thus, this study focuses on developing physical computing educational materials intended for university

students who have some experiences with using computers [1]. We named this the "AI Builder Learning Kit". The rationale for incorporating physical computing is that it has been used as an accessible teaching method for beginners in programming education [6]. Physical computing allows students to perceive errors through physical movements. By utilizing physical computing materials, the authors have found that students easily acquire of AI literacy. In the evaluation experiment, students will study using the developed teaching materials and verify the effectiveness of the materials by comparing them with learning using textbooks alone.

II. SUGGESTED AI LEARNING MATERIALS "AI BUILDER LEARNING KIT"

In this study, we have carried out the following steps for the development of our educational materials.

1. Establish learning standards for AI literacy based on literatures.
2. Develop physical computing educational materials modeled on autonomous driving, based on the established learning standards.
3. Verify whether the materials can be used for learning.
4. After verification, conduct an evaluation experiment of the proposed materials to assess their effectiveness in improving awareness and knowledge related to AI. (Planned for the future)

A. Developing Learning Standards for AI Literacy

In this study, we have established learning standards necessary for acquiring AI literacy, aiming to experiential learning for everyday use of AI. To define the learning standards, we have investigated models of curricula recommended by consortia dedicated to strengthening education in mathematics, data science, and AI, as well as the G exam, a Japanese certification that tests knowledge of machine learning [7] [8]. This approach ensures that the learning standards cover essential aspects of AI and responsibilities using AI technologies in daily life.

The reference literature [7] for this material is based on the "AI Strategy 2019" established by the Japanese government in 2019. Therefore, this material follows Japan's traditional educational methods.

1) Mathematics/Data Science/AI Model Curriculum

We employed the "Mathematics, Data Science, & AI (Literacy Level) Model Curriculum – Cultivating Data Thinking" for setting the AI learning standards [7]. The

learning objective of this curriculum is defined as "to proactively acquire the foundational proficiency necessary to proficiently apply mathematics, data science, and AI in daily life, work, and other scenarios." The emphasis is on "the capability to make appropriate, human-centered decisions." The fundamental approach includes "a focus on teaching the 'joy' and 'significance' of engaging with and learning about mathematics, data science, and AI." The curriculum orderly presents learning items and is systematically structured as shown in Table I. Within this structure, the areas related to AI learning include sections 1-3, 1-4, 1-5, 1-6, and 3-1.

In the first section, "Utilization of Data & AI in Society," the focus is mainly on AI knowledge and application, presenting skill sets for specialized AI and general-purpose AI, among others. The second section "Data Literacy" discusses how to handle data, but it hardly mentions AI, hence we do not focus on this study. The third section, "Considerations in the Utilization of Data & AI," suggests covering negative examples of AI utilization and data ethics, among other topics.

TABLE I. STRUCTURE OF AI LITERACY LEVEL MODEL CURRICULUM

1. Introduction <i>Utilization of data and AI in society</i>	1-1 Changes occurring in society
	1-2 Data used in society
	1-3 Application areas of data and AI
	1-4 Technology for data/AI utilization
	1-5 Fields of data/AI utilization
	1-6 Latest trends in data and AI utilization
2. Basic <i>Data literacy</i>	2-1 Read the data
	2-2 Explain the data
	2-3 How to use data
3. Knowledge Considerations in the Utilization of Data & AI	3-1 Points to note when handling data and AI
	3-2 Points to note when protecting data

2) *DLA Deep Learning For GENERAL*

In order to survey the required knowledge of machine learning, we investigated the official textbook for the Deep Learning G Certification, "Deep Learning G Certification Official Textbook 2nd Edition," which is structured according to the syllabus of the qualification examination. The official textbook is comprised of seven chapters, with contents as follows:

1. What is Artificial Intelligence (AI)?
2. Trends Surrounding Artificial Intelligence.
3. Issues in the Field of Artificial Intelligence.
4. Specific Methods of Machine Learning.
5. Overview of Deep Learning.
6. Methods of Deep Learning.
7. Toward the Social Implementation of Deep Learning.

We have focused on Chapters 1, 2, and 7. Chapter 1 discusses the nature of AI, including its history and classification, and explains the differences between machine

learning and deep learning at various levels. Chapter 2 addresses the trends in AI, emphasizing the history and relationship of machine learning and deep learning research. It particularly notes that desirable results can be achieved through accumulating data in machine learning and explains the mechanisms of machine learning and deep learning differ, and how they are different. Chapter 7 covers methods and considerations for utilizing AI towards social implementation. The chapter also discusses how to handle data, including the quality of datasets. It emphasizes how to eliminate bias, and how to process and to analyze data fairly, and how to learn regularity from data.

The mathematical, data science, and AI (literacy level) curriculum explicitly focuses on data science, primarily statistics, with AI employed as a means within this context. The fundamental approach emphasizes the 'fun' and 'significance' of learning, which motivates students to engage actively and enjoyably with AI.

From the 'G certification' perspective, the curriculum is based on statistical operations that can be learned in mathematics, data science, and AI, highlighting how AI can be utilized in the real world. It includes understanding what AI is, its mechanisms, and foundational knowledge, while also emphasizing the importance of 'how data can be applied.' The recurrent themes of data quantity and quality are considered the most crucial knowledge for learning AI.

Based on the observation of Chapters 1 and 2, we have formulated the foundational learning criteria and perspectives on AI, which are presented in Table II.

TABLE II. FOUNDATIONAL LEARNING CRITERIA AND PERSPECTIVES ON AI

	Learning Criteria & Perspectives	Points of Understanding
A	Generality & Specificity	Specializes in performing certain tasks (e.g., image & voice recognition)
B	Learning & Training Data	The operation of AI is indispensable for learning data, with the quality and quantity of data being crucial
C	Validity of Inference Results	The quantity and quality of training data can affect achieving the desired results

The rationales for formulating each perspective are as follows:

- A. From the perspective of "human-centered" importance in mathematics, data science, and AI, it is necessary to learn about what AI can and cannot do.
- B. As handled in prior research and teaching practices, approaches to collecting learning data for image recognition, the emphasis on the consciousness of statistical work for data utilization in mathematics, data science, and AI, and the G Certification's point on the necessity of processing, analyzing, and learning the training data for AI's social implementation are reasons for this perspective.
- C. The G Certification mentions that desirable results can be achieved depending on the quantity and quality of data, underlining the necessity to understand that the desired outcomes may not always be attainable depending on the data.

B. Physical Computing Teaching Materials

We developed educational materials for experiential physical computing that allow students for comprehensive learning of the established learning criteria and perspectives. The goal of these materials is to motivate AI learning and enable active learning. As shown in Figure 2, the teaching materials and PCs are connected via Wi-Fi. Students learn machine learning while checking the camera input on the PC.

1) Specifications of the Educational Material

In this research, we developed a mobile robot-like educational material that can recognize signs through image recognition using the Jetson Nano B01, a single-board computer for AI learning released by NVIDIA [9]. It controls the robot according to the meaning of the signs. Figure 1 shows the developed robot-like educational material, Figure 2 shows the hardware configuration. The robot-like educational material communicates with the server using wireless network so that it executes the AI learning model.

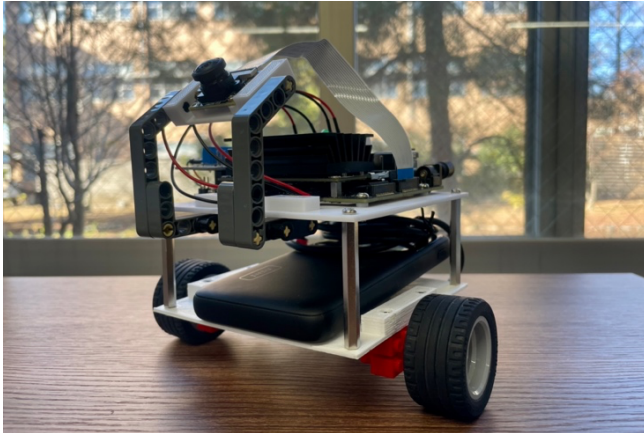


Figure 1. Developed physical computing teaching materials.

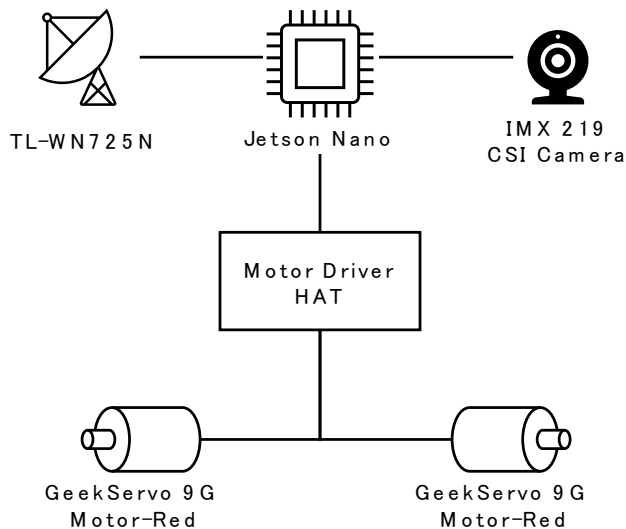


Figure 2. Hardware configuration diagram.

The software configuration is as follows. We employed the JetPack 4.6 platform for Jetson, published by NVIDIA, and Docker containers used by the NVIDIA Deep Learning Institute. They allow operators to access Jupyter Lab via a browser. Additionally, we used pre-installed PyTorch, which is the machine learning library used in this container.

The specifications of the educational materials are shown in Table III, and the components of the materials are listed in Table IV.

TABLE III. SPECIFICATIONS OF THE DEVELOPED EDUCATIONAL MATERIALS

Specification	Details
Dimensions	170mm (W) x 110mm (D) x 150mm (H)
Power Source	Lithium-ion battery
Continuous Operation Time	4 hours
Charging Method	USB charging via USB Type-C

TABLE IV. LIST OF COMPONENTS FOR THE EDUCATIONAL MATERIALS

Category	Component Name	Quantity
Controller	Jetson Nano B01	1
Drive Motor	GeekServo 9G Motor-Red	2
Motor Driver I2C Interface	WaveShare 15364 Motor Driver HAT for Raspberry Pi	1
Camera	Yahboom IMX219 160-degree CSI Camera	1
Wi-Fi Module	TP-Link TL-WN725N	1
Battery	INIU POWERBANK BI-B6	1
Tires	LEGO 4184286	2
Wheels	LEGO 4297210	2
Caster	TAMIYA No. 144 Ball Caster	1

2) Overview of the Robot-like Educational Material

This robot-like educational material recognizes signs and proceeds according to the meaning of those signs through image classification using Convolutional Neural Networks (CNN). The material developed for this occasion classifies two classes (background and signs). We conducted experiments utilizing a sign indicating a speed limit of 10 km/h to slow down the operational speed of the material. Additionally, as an advanced application, there is a program that classifies six classes. Table V shows the recognized objects and corresponding actions.

TABLE V. CORRESPONDENCE TABLE OF RECOGNIZED OBJECTS AND ACTIONS

Recognition Objects	Actions
Background	Normal operation
Speed limit 10km	Operating speed 10
Speed limit 30km	Operating speed 30
Stop	Pause (1 second)
No entry	allowed End of operation
People	Stop until no more people are classified

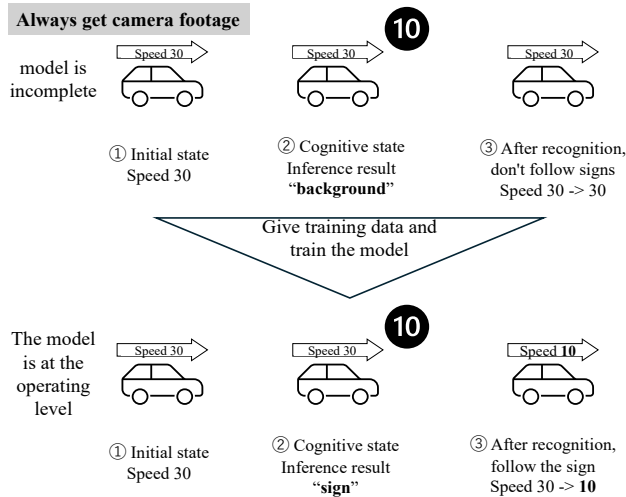


Figure 3. Operational image of the educational materials.

Figure 3 shows the conceptual image of the operations of the robot-like educational material.

The educational materials are structured around five processes based on the perspectives shown in Table I. By sequentially implementing these processes, students can experience and study image classification AI. The established processes and the corresponding learnable perspectives they cover are shown in Table VI.

TABLE VI. LEARNING PROCESSES AND PERSPECTIVES IN THE EDUCATIONAL MATERIALS

Step	Content	Perspective
1	Prepare the learning data	B
2	Define the model	A
3	Train the model	B
4	Test the model	C
5	Adjust the data based on results	B

III. VERIFICATION OF EDUCATIONAL MATERIALS

Students can perform a series of AI learning activities by accessing Jupyter Lab via a browser. He or she must follow the steps in Table VI.

First, the student performs the step 1 through 4. Of course, the robot-like educational material cannot classify signs and executes incorrect actions. Then, the student proceeds to step 5 to adjust training data and the frequency of training sessions. Then he or she iterates the procedure steps 3, 4 and 5 until the robot-like educational material achieves the accurate inference. Figure 4 and 5 shows the experiments of this procedure.

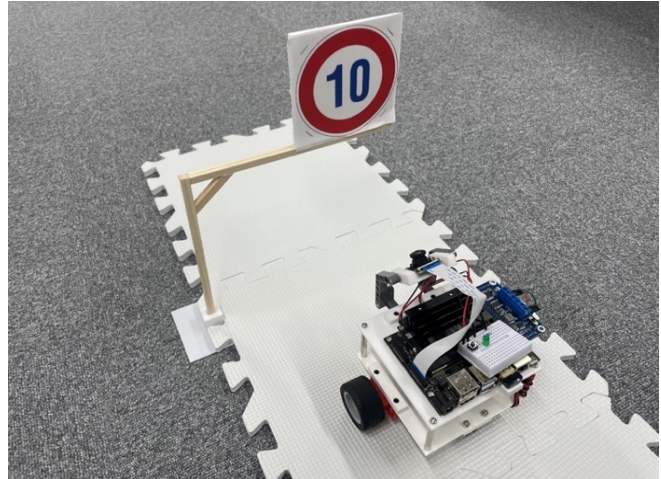


Figure 4. Operational aspect of the proposed educational material.



Figure 5. Learning data collection screen.

In step 3, the robot-like educational material collects learning data through a camera mounted on it.

For step 5, adjusting the training data, students individually modify the learning data and model training. Adjusting the learning data involves increasing the data volume based on the operational results. For the model training, we increase the number of learning iterations until the loss is stabilized, since the system presents the number of epochs and the loss graph. After adjustments, students check the accuracy of the model through the operation of the robot-like educational material. Figure 6 shows the control panel of the system. The students can adjust the data and learning.

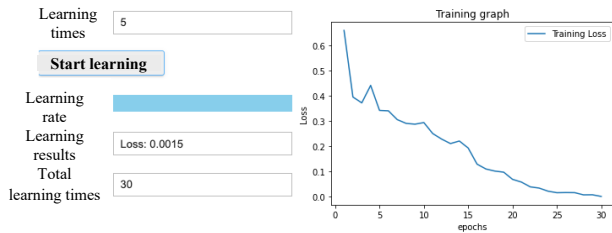


Figure 6. Adjustment of model training.

This teaching material is built based on the official PyTorch tutorial. Students can observe how the model training progresses using the control panel shown in Figure 5 without programming.

A. Verification of Learning Standards

The proposed educational materials incorporate image recognition. We must investigate whether the students can comprehend the learning standards through the experiences of image recognition students. We made students engage in an AI experience focused on character recognition (image recognition) aligned with the learning standards to assess their adequacy.

Following this experience, Table VII shows the collected responses on the comprehensibility of each learning standard. We have confirmed that the students deeply comprehend the learning standards we proposed by studying image recognition.

TABLE VII. VALIDITY OF LEARNING CRITERIA FOR IMAGE RECOGNITION N=3

Learning Criteria	Understood	Not Understood
A	3	0
B	2	1
C	3	0

B. Supplementary Textual Educational Materials

To supplement the knowledge that cannot be fully covered through the learning flow and experience of the developed educational materials, a roadmap-style text-based educational resource was created. Additionally, for the evaluation experiment, a web application was developed that displays the text and records the viewing time for each page.

In this study, two different types of text-based educational materials were developed for the evaluation experiment. The first is a text that serves as a guideline for using the physical computing materials. The second is a standalone text that allows students to complete their studies without using physical computing materials.

The text-based educational material consists of 19 pages, divided into two parts. The first part explains the basic knowledge based on the G certification. The second part provides a roadmap for experiential learning using the physical computing materials. Table VIII shows the correspondence between the content of each page of the text-

based material for using the physical computing materials and the relevant perspectives.

TABLE VIII. CONTENT AND PERSPECTIVES OF THE PHYSICAL COMPUTING-BASED TEXT EDUCATIONAL MATERIAL

Page	Content	Perspective
1, 2, 7	Cover, Section Cover Pages	-
3	What is AI?	A, B, C
4	AI Excels at Specific Tasks	A
5	AI Learning Process	B
6	The Four Levels of AI	A, C
8, 9	Introduction of the Materials, Learning Objectives, Device Operation	-
10	Executing the Setup Program Cells	-
11	How to Capture Learning Data	B
12	Preparing Learning Data (Data Limitations)	B
13	Defining the AI Model	-
14	Model Training (Specifying Number of Iterations), First Round	B
15	Verifying Operation, First Round	B, C
16	Model Training (Removing Iteration Limitations), Second Round	B
17	Verifying Operation, Second Round	A, C
18	Enriching Learning Data, Model Training	B
19	Verifying Operation, Third Round	A, C

When using educational materials in Jupyter Lab, it is important to ensure that students can follow textual instructions and terminology to accurately execute operations. Each page of the text contains themes to be learned and the content of the cells to be executed.

Furthermore, an application has been developed using PHP and JavaScript to record the viewing time of each text page and the execution time of each cell, viewable as images online. The application allows users to switch between text pages using 'Next' and 'Back' buttons, capturing the time spent on each page. The 'I have learned and executed' button records the time from when the page is displayed to when the button is pressed. The application for viewing the text is shown in Figure 7. The top of Figure 7 contains a bar graph that allows users to check their learning progress, the center contains a text display area, and the bottom contains an operation area.

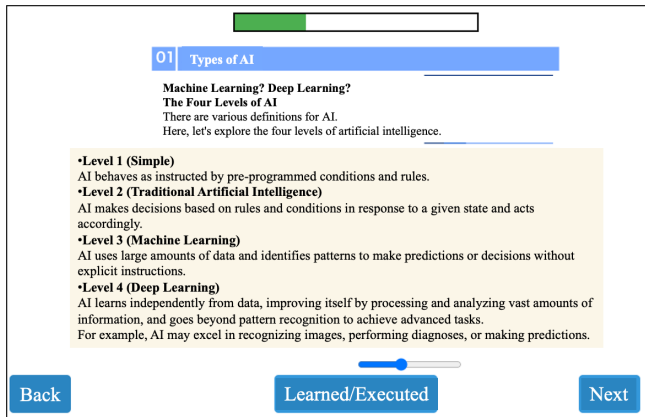


Figure 7. Application for viewing textbooks

IV. EVALUATION EXPERIMENT

To verify whether the established learning standards could be met through learning with physical computing materials, an evaluation experiment was conducted with 17 university science students, divided into an experimental group (8 students) and a control group (9 students).

A. Experimental Procedure

The evaluation experiment proceeded as follows:

1. Conduct a preliminary survey titled "Survey on Awareness about AI" (hereafter referred to as the "Awareness Survey") with the 17 participants.
2. The control group worked on materials based on the official PyTorch tutorials corresponding to section 3.B of the text, while the experimental group used physical computing materials along with the text from section 3.B.
3. Conduct a post-survey titled "Knowledge Survey about AI" (hereafter referred to as the "Knowledge Survey").

The Awareness Survey, designed based on prior studies and practices related to AI awareness, aims to determine if the educational materials effectively motivate and enhance awareness. The items of the Awareness Survey are listed in Table IX. The evaluation scale for Question 4 consists of four levels: 1. Don't know, 2. Have heard of it, 3. Know it, and 4. Can explain it.

The Knowledge Survey checks whether participants understand each item based on learning standards and perspectives, with questions set from knowledge and application skills necessary for G certification and learning standards. The items of the Knowledge Survey are shown in Table X. The correct answer criteria for Question 5 are met when the steps are presented in the following order: 1. Prepare the data, 2. Build the model, 3. Train the model, and 4. Test and verify the completed model.

TABLE IX. AWARENESS SURVEY ITEMS

No.	Item	Collection/Evaluation Method
1	Choose the closest representation of your future vision of AI	Four-point scale
2	How do you feel about the improvement of life with AI development?	Four-point scale
3	What are your feelings towards AI?	Four-point scale
4	Choose the description that best matches your understanding of AI	Four-point scale
4-1	How is AI created?	
4-2	Why can AI recognize images or generate text?	
4-3	What can AI do?	
4-4	The difference between AI and robots	

TABLE X. KNOWLEDGE SURVEY ITEMS

No.	Item	Collection/Evaluation Method
1	Choose the closest term related to AI from the options provided	Four-point scale
1-1	Machine learning	
1-2	Deep learning	
1-3	Dataset	
1-4	Image recognition	
2	Select whether AI can perform the following tasks	True/False
2-1	Recognize a face and identify the individual	
2-2	Diagnose diseases accurately	
2-3	Determine the cause of a machine breakdown	
2-4	Explain information about an event that occurred yesterday	
3	Select all correct processes necessary for developing an image recognition AI model	True/False
4	Given a dog image was classified as a cat despite sufficient training, what is the most likely cause, assuming correct classification of other dogs?	True/False
5	Rearrange the steps to create an AI	True/False

B. Experimental Results

The effectiveness of enhancing awareness was evaluated by comparing the changes in the Awareness Survey results before and after the experiment using a two-tailed t-test. The results are shown in Table XI.

TABLE XI. RESULTS OF THE AWARENESS SURVEY BEFORE AND AFTER IMPLEMENTING EDUCATIONAL MATERIALS

Survey Item	Pre-Survey		Post-Survey		t-test	
	M	SD	M	SD	t-value	
Vision of AI	4.00	0	4.00	0		n.s.
Support for AI Development	3.63	0.48	3.63	0.48		n.s.
Support for AI Utilization	3.50	0.50	3.88	0.33	-2.05	n.s.
Understanding of AI Mechanisms	2.00	0.71	2.88	0.33	-2.97	*
Principles of AI	1.88	0.78	3.25	0.43	-4.25	**
Applications of AI	2.75	0.83	3.38	0.48	-2.38	*
AI vs. Robots	2.13	0.93	3.13	0.60	-3.06	*

n=8, *: $p < .05$, **: $p < .01$

Improvements were observed in most survey items, except those already had high evaluations before the implementation. Notably, the understanding of AI principles significantly increased, as indicated by the statistics ($t(7) = -4.25$, $p < .01$). This suggests that the educational materials effectively enhanced comprehension of AI principles.

Next, to verify the validity of the materials, the knowledge survey checked terminology on a four-point scale, while other items were scored as 1 for correct and 0 for incorrect answers. Changes between the experimental and control groups were evaluated using a two-tailed t-test. The results are shown in Table XII.

TABLE XII. RESULTS OF THE KNOWLEDGE SURVEY AFTER IMPLEMENTING EDUCATIONAL MATERIALS

Quiz Category	Experimental Group n=8		Control Group n=9		t-test	
	M	SD	M	SD	t-value	
Terminology	3.16	0.51	3.53	0.55	-2.85	**
Applications of AI	0.84	0.36	0.58	0.49	2.46	*
Applied Problems	0.69	0.46	0.50	0.50	1.10	n.s.
Procedures Overall	0.50	0.50	0.67	0.47	-0.66	n.s.
Correct Answer Rate	0.75	0.12	0.52	0.21	2.89	*

*: $p < .05$, **: $p < .01$

In the terminology section, the control group tended to score higher. However, the experimental group showed higher average scores in practical applications, and although no significant difference was found in application problems, the correct answer rate was higher. Since significant differences were observed in problem-solving rates, the educational materials are considered effective for enhancing practical AI knowledge and as introductory materials for AI learning.

V. RELATED WORK

Scratch is a well-known programming learning material. Scratch is extensible and now it includes materials focused on machine learning. An example is ML2Scratch, which enables image classification using MobileNet through TensorFlow.js [10]. Furthermore, based on this study, researchers have developed another extension that allows for the learning of advanced deep learning techniques such as transfer learning [11].

Google's Teachable Machine is a web-based tool that easily allows to create machine learning models [12]. It supports to create models of three categories, i.e., image, sound, and pose. We can create and export a TensorFlow.js models through collecting learning data directly on the site by taking pictures or recording sounds, and with the press of a training button. A research at the University of Potsdam has shown that utilizing physical computing educational materials promotes not only intrinsic motivation but also creative and constructive learning [13].

Felix Hu and colleagues developed a tangible programming game called "Strawbies" for children aged 5 to 10 years [14]. The game involves programming with wooden tiles, which are not square but specially shaped to prevent incorrect connections. Although this design reduces the freedom of programming, it offers the benefit of allowing users to intuitively understand whether a connection is possible or not. Aditya Mehrotra and his team implemented robot programming classes where students rearranged printed program blocks, and they evaluated several methods [15]. However, the purpose of their study was to evaluate the methods themselves, and they did not adjust the instructional content in real-time based on students' progress to ensure knowledge retention. Kato and others developed and evaluated a system that collects and analyzes students' programming progress, providing this information to assist instructors in efficiently guiding students [16]. However, since this analysis focuses on programming languages, it cannot be directly applied to tangible educational materials.

Regarding these studies, materials using Scratch are web-based, resulting in outcomes being displayed on the screen, akin to the initial experiences of text display in programming learning. Teachable Machine specializes in model creation. While exporting models allows for a broad range of learning opportunities, advancing in applied learning requires prior knowledge of the application areas. A commonality among these materials is their use of transfer learning, which tends to produce relatively accurate results. Although it is easy to obtain results from machine learning through these examples, they do not help for deepening knowledge. We address this problem.

Table XIII shows the previous cases and the characteristics of the authors.

TABLE XIII. COMPARISON WITH OTHER STUDIES

<i>Name</i>	<i>Physical Computing</i>	<i>AI Learning</i>	<i>Multi-Student</i>	<i>Analysis on Class</i>	<i>Analysis after Class</i>
AI Builder Learning Kit	+	+			+
ML2Scratch [10]		+	+		
Teachable Machine [12]		+	+		
Strawbies [14]	+				
PaPL [15]	+		+		+
Katos' System [16]			+	+	+

VI. CONCLUSION AND FUTURE WORK

In this study, we established fundamental learning standards and perspectives for learning the basic mechanisms of AI and developed educational materials that align with these standards. As a result, students' awareness of various aspects of AI improved, and their understanding of its mechanisms and principles increased. These outcomes suggest that the developed materials can enhance both the understanding of basic AI mechanisms and literacy in AI-related awareness.

Future challenges include making it easier to learn terminology that was not fully covered by the current materials and improving the materials to incorporate generative technologies rather than just classification.

ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI, Grant Number JP24K06237.

REFERENCES

- [1] T. Kato and Y. Chino, "Development of AI learning materials using physical computing," The First International Conference on IoT-AI, IARIA, 2024.
- [2] S. Noy and W. Zhang, "Experimental evidence on the productivity effects of generative artificial intelligence," *Science* 381.6654: pp. 187–192, 2023.
- [3] Cabinet Office: AI Strategy 2019. [Online]. Available from: <https://www8.cao.go.jp/cstp/ai/aistrategy2019.pdf> 2023.12.23
- [4] N. Takezawa, T. Yamamoto and H. Koura, "Proposal and Evaluation of a Teaching Process for Learning Image Recognition Technology Using Artificial Intelligence through Programming," *Japan Society of Educational Information*, vol. 38, no. 1, pp. 37–48, 2022.
- [5] N. Mukaïda, "Examination of Classes to Think About How to Face AI in AI Literacy Education," *Japan Educational Research Society of the AI era*, vol. 5, pp. 9–15, 2022.
- [6] P. Mareen and R. Ralf, "Impact of Physical Computing on Learner Motivation," *Koli Calling '18*, pp. 1–10, 2018.
- [7] Consortium for Strengthening Mathematics and Data Science Education in Japan, "Mathematics/Data Science/AI (Literacy Level) Model Curriculum -Cultivating Data Thinking-, " [Online]. Available from: http://www.mi.u-tokyo.ac.jp/consortium/pdf/model_literacy.pdf 2024.01.04
- [8] Japan Deep Learning Association, "Deep Learning G Test Official Text, " 2nd edition, 403p, 2021.
- [9] NVIDIA, "Jetson Nano developer kit, " [Online]. Available from: <https://www.nvidia.com/ja-jp/autonomous-machines/embedded-systems/jetson-nano-developer-kit/> 2024.01.04
- [10] J. Ishihara, "Introducing an extension that makes machine learning available from Scratch," *Interface*, vol. 47, no. 5, CQ Publishing, pp. 148-152, 2021.
- [11] T. Yagi and T. Yamaguchi, "Building a machine learning experience system for beginners using Scratch, " *Bulletin of Edogawa University*, vol. 30, pp. 473-485, 2020.
- [12] Google: Teachable Machine [Online]. Available from: <https://teachablemachine.withgoogle.com/> 2024.01.17
- [13] M. Przybylla and R. Romeike, "Impact of Physical Computing on Learner Motivation, " *Koli Calling 2018*, pp. 1-10, 2018.
- [14] F. Hu, A. Zekelman, M. Horn and F. Judd, "Strawbies: explorations in tangible programming," *IDC '15: Proceedings of the 14th International Conference on Interaction Design and Children*, pp. 410-413, Boston Massachusetts, United States of America, June, 2015.
- [15] A. Mehrotra, C. Giang, N. Duruz, J. Dedelley, A. Mussati, M. Skweres and F. Mondada, "Introducing a Paper-Based Programming Language for Computing Education in Classrooms," *ITiCSE '20: Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education*, pp. 180-186, New York, United States of America, June, 2020.
- [16] T. Kato, Y. Kambayashi, Y. Terawaki and Y. Kodama, "Estimating Grades from Students' Behaviors in Programming Exercises Using Deep Learning," *International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 1115-1119, IEEE, 2017.