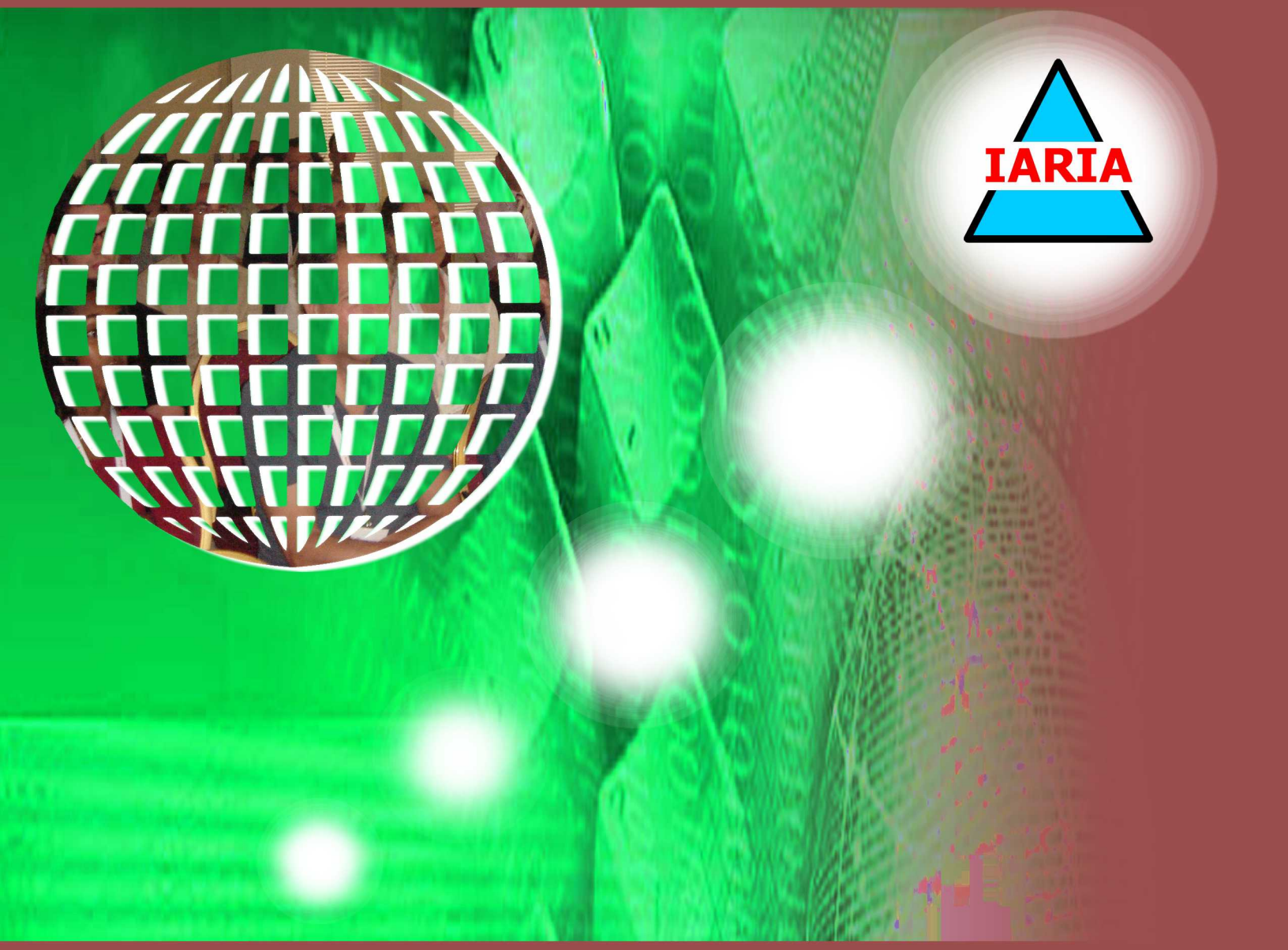
International Journal on Advances in Life Sciences



2024 vol. 16 nr. 3&4

The *International Journal on Advances in Life Sciences* is published by IARIA.
ISSN: 1942-2660
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Life Sciences
Volume 16, Number 3 & 4, 2024

International Journal on Advances in Life Sciences
Volume 16, Numbers 3 & 4, 2024

CONTENTS

Learning Biomechanics Applied to Orthodontics: Interest and Characteristics of an Innovative Simulation Device

Aurelie Mailloux
Reims Hospital, URCA
Reims, France
Email: aurelie.mailloux@univ-reims.fr
2LPN UR 7489

***Abstract*—Healthcare simulation devices offer the opportunity to improve practitioners' knowledge, skills and behaviors. The biomechanical simulation tools currently used in orthodontics are technically and pedagogically limited. The simulation tool (through dynamic visualization) could enable students to better understand the biomechanical principles of tooth movement, both in theoretical courses and clinical practice. The need to develop an innovative simulation device in this field is indeed shared by orthodontics students. The aim of this article is to identify (i) the interest in developing a innovative digital simulation device in this field, (ii) the learning objectives, (iii) the technical aspects, (iv) the constraints for designing an innovative device.**

Keywords*—*Training; Simulation device; Orthodontics; Biomechanics

I. Introduction

Previous studies on orthodontic students' training expectations revealed (i) the limitations of current biomechanical simulation tools, (ii) the need to develop an innovative simulation device in this field [1][2]. Section I provides background on (i) biomechanical concepts applied to orthodontics (ii) current simulation devices in the field of orthodontics and (iii) their pedagogical results. Section II describes the data collected to identify priority application areas for the development of a biomechanical simulation tool. Sections III to V describe the pedagogical objectives, as well as elements of further studies that should be discussed and carried out on the subject.

A. Context

1) Biomechanics applied to orthodontics: Whatever the appliance used, orthodontic movements are based on biomechanical concepts. To understand tooth movement, it is necessary to represent the equivalence of the system of forces at the tooth's center of resistance (CR) [3]. The CR is a theoretical point on the tooth. When a force is applied to it, the tooth is displaced in translation (i.e., without causing a version of the tooth crown). The CR is located on the long axis of the tooth. The location of the CR depends on the height of the alveolar bone, the length of the root and the number of roots (Figure 1). In orthodontics, forces cannot pass directly through the CR (i.e., forces are applied to the orthodontic bracket, bonded to the crown). The distance between the CR and the orthodontic bracket is therefore variable.

Tooth translation involves moving the tooth along the occlusal plane without altering the orientation of the major axis. As shown in Figure 1, this movement is impossible to achieve

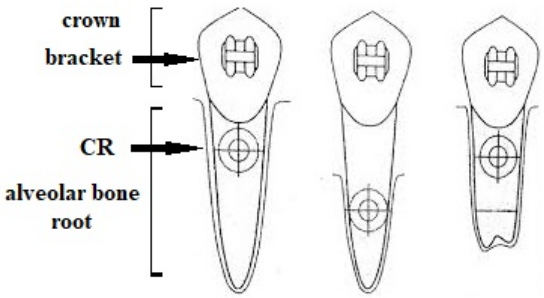


Figure 1. The different locations of a canine CR depend on the length of the root and the height of the alveolar bone. The distance between the orthodontic bracket and the CR modifies the forces system and tooth movement

using a simple linear force on the bracket. As shown in Figure 2, the application of a simple linear force on the orthodontic bracket creates an uncontrolled rotational movement of the crown and tooth root around the tooth's center of rotation (close to the CR).

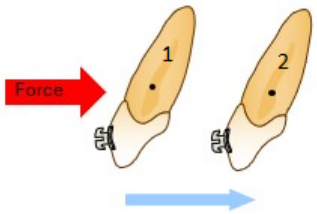


Figure 2. A simple force applied on the center of resistance implied a movement of translation (parallel to the long axis of the tooth, from position 1 to position 2)

So, as Figure 3 shows, there is a rotational and/or version movement (i.e., moment) in addition to the linear movement (i.e., called the translation movement). The force system depends on (i) the initial clinical situation, (ii) the chosen chosen appliance and (iii) the orthodontic technique (e.g., segmented or continuous). Learning biomechanics can be challenging, as the concepts are hard to grasp in a static context [4].

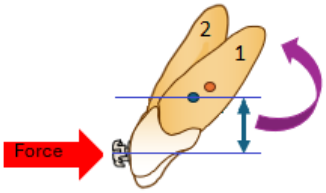


Figure 3. A simple force applied on the tooth orthodontic bracket (far from the center of resistance) implied a non-controlled tooth movement, from position 1 to position 2

2) Current simulation devices for orthodontists: Several simulation devices dedicated to orthodontic practitioners have recently been developed in the following formats: 3D, immersive learning, haptics, augmented reality, serious games. The simulation devices identified have been implemented in various fields such as education, covering diagnosis and treatment planning, orthodontic bracket positioning, orthodontic procedures, facial marking, orthodontic aligners and cephalometric tracing. Most studies dealing with biomechanical simulation focus on the development process without evaluating the results [5]. Most simulation devices were designed specifically to facilitate orthodontic treatment planning and procedures. They incorporate biomechanical concepts to anticipate tooth movements according to the chosen archwires (size, section and materials). However, these systems seem to be limited. For example, it is not possible to simulate anchorage, such as mini-screws or extra-oral forces, or elastomeric chain (even though these devices are part of current daily practice). Orthodontic anchorage is defined as a means of resisting the movement of one or more teeth, using different techniques or tools. Anchorage is indeed an important consideration in treatment planning, as unplanned or unwanted tooth movement can have disastrous consequences. Furthermore, the purpose of these simulation devices is not pedagogical: they have been developed for private orthodontic practitioners (i.e., not for orthodontic students).

In the orthodontics field, clinical skills are currently taught through demonstrations on patients (by trial and error). The use of technology is currently limited and poorly designed [6]. Orthodontics novices currently learn biomechanics (i.e., theoretical knowledge and practical skills) with static schemes or using an occlusor, made of metal teeth embedded in a sticky wax, as Figure 4 illustrates. Sticky wax is a mixture that dissolves in water at the temperature of 60–65°C (i.e., by heating, it allows dental movements). However, this system (i) does not faithfully reproduce bone remodeling and anchorage (e.g., mini screw), (ii) does not make it possible to visualize root displacements, nor the successive stages from the initial to the final clinical situation.

The aim of this paper is to list the characteristics of an "ideal" simulation device for learning orthodontic biomechanics and therefore choosing the mechanics best suited to their patients' malocclusion.

Figure 4. Orthodontic occlusor

II. Digital devices dedicated to learning orthodontics

A. Pedagogical outcomes

Orthodontics is a discipline that requires the acquisition of theoretical knowledge in various fields (such as head and neck anatomy, maxillary growth and development, physiology and biomechanics of tooth movement) and a number of manual skills (e.g., archwire bending, bracket positioning, dental stripping). The practice of orthodontics requires knowledge and skills in a variety of areas. Several experiments have demonstrated the effectiveness of digital simulations in terms of knowledge acquisition and improved reaction levels. These results are consistent with the impact of digital simulations, including virtual and augmented reality, in other areas of dental education [7]. For example, 92 percent of students understand dental anatomy better using virtual learning than using a traditional written document [8]. Virtual learning improves X-ray detection of bone lesions compared with traditional courses [9]. More generally, virtual learning could enhance theoretical knowledge.

Concerning the improvement of medical gestures, gesture recognition based on wearable sensors has developed in the health and dental sector in recent years [10]. Dexterity improvement is difficult to assess, and there is a lack of data in this area [11]. No orthodontic digital tool dedicated to learning biomechanics and/or mastering movements has been found in the literature.

The collected data confirmed that digital and simulation tools could be efficient in terms of learning. This simulation device could therefore help orthodontic students gain expertise (i.e., to enhance theoretical knowledge, and practical skills).

B. Interest of an innovative biomechanics' simulation device

Simulation devices in the health field aim at securing patient care. They are based on the concepts of (i) "never the first time on a patient" and also (ii) "mastering gestures before treating patients". Simulation allows training in semi-real conditions, which makes the learner more involved than during lectures (i.e., with a top-down teacher-learner scheme). A meta-analysis conducted in 2011 highlighted the possibility of improving practitioners' knowledge, skills and behaviors through health simulation [12]. Appropriate substitutes for clinical practice should be considered to ensure

that students treat patients without making errors damaging to their orthodontic treatment. Moreover, animations could help practitioners understand complex dynamic processes in a simple and realistic way [4]. According to a collective of orthodontic experts, simulation tool and video could be very helpful to enhance students' understanding of bio mechanical principles of tooth movement (such as forces control and the moment/force ratio). The visualization of tooth movement (depending on the chosen appliance) could achieve a clear understanding of how tooth's crown and roots move inside the bone. This could also allow students to predict the treatment outcomes (based on the treatment plan and the chosen system of force).

The future learning device should enhance both biomechanical theoretical knowledge and manual skills. Thus enabling orthodontic students to acquire clinical expertise.

C. Technological aspects

Technological advances (e.g., imagery/ 3D radiography and computer image processing) enable to obtain specific anatomical models of a patient, meshable and usable by finite element software [13]. Thanks to the monitoring, it is possible to correlate finite element analyses with clinically observed movements. However, the finite element approach still has some limitations:

1- Long-term tooth movement cannot be predicted from the initial force system

2- Tooth movement depends on (i) the characteristics of the patient (e.g., drugs, dental morphology, alveolar bone, masticatory forces, tongue), (ii) the force system (e.g., continuous or segmented archwire, alloy, friction)

In addition, computational modelling remains complex and time-consuming [14]. In the literature, studies conducted on finite elements applied to orthodontics aimed at improving:

1- The treatment planning optimization and individualization (e.g., choice of the archwire in accordance with the clinical situation)

2- The anticipation of iatrogenic damages (i.e., caused by orthodontic treatments)

3- The accuracy of the forecasts of the treatment results

Thus, the implementation of an optimal orthodontic force system model that meets all these requirements is challenging. New studies are underway to improve digital modelling precision. Some studies have already combined experimentation (i.e., to quantify forces and moments, using test beds) and digital modelling [14]. Furthermore, the scan of different stages of orthodontic treatments could improve the management of similar clinical situations (i.e., machine learning is already used for treatment planning by aligners). This approach is similar to what has been assumed in medical practice: records of the best clinical decisions made by thousands of professionals should be leveraged to improve patient care and practitioners training [15]. Along with these, we believe that from a learning and training perspective, the current technologies are sufficient to design new simulation-based learning activities in biomechanics. These should allow orthodontic students to (i) improve their manual skills, (ii) anticipate the side effects of the appliances, (iii) be able to choose the most suitable device(s) according to the initial patient clinical situation.

III. Priority fields for a future simulation device

The mastery of biomechanics applied to orthodontics requires both theoretical knowledge and manual skills in various areas, such as archwire bending or bracket positioning. From the literature, we have carried out the following classification concerning the theoretical knowledge necessary to understand biomechanics. It summarizes the sub-dimensions of biomechanics applied to orthodontics:

1- Physiology of tooth movement (physiological periodontal and bone response related to orthodontic strength...)

2- Tooth movement (the three orders, theory, indications)

3- Force systems (moment/force, equilibrium...)

4- Anchorage (anchorage and its control, mini-implants...)

5- Fixed devices (treatment mechanics, vectors, forces, moments applied, arches deformations and constraints)

6- Biomaterials related to tooth movement (biomaterials and production of orthodontic forces...)

7-Removable Device (interception and prevention...)

8-Factors affecting tooth movement (patient factor, growth)

9-Iatrogenic effect of tooth movement

We therefore interviewed the Reims Hospital students (N=6) to identify the priority fields for developing simulation tools, among this classification. They have considered the following sections, as priorities: force system (9 points), anchorage (7 points) and fixed appliance (6 points). Points were assigned to each response based on their rank of importance. This survey should be extended to a wider pool of students in orthodontics, in order to ensure that this order suits them.

IV. Educational Goals

According to the identified priority fields and the current biomechanics learning objectives, an effective simulation device should allow orthodontic students to:

- scan a patient's malocclusion and virtually position the brackets on the tooth crown.

- improve their manual skills: (i) scan archwires bent by students on a real-life occlusor, and integrate them in the simulator to evaluate and visualize the dental movements they generate, (ii) compare the ideal archwire with the one bended by the student.

- visualize the dental movements according to the clinical initial situation and the chosen fixed appliance. The dental displacements should be split into successive steps (i.e., from the initial to the final situation) by showing and quantify the forces and the moments on each tooth (i.e., including the visualization of dental root movement)

- evaluate the probability with which the movement will occur, according to the therapeutic choice and the treatment objectives (such as control of the mandibular incisor axis, finishing with a dental Angle's class I, preparation for maxillofacial surgery)

- send alert messages according to the chosen therapeutic and treatment objectives (e.g., insufficient anchoring to obtain the attempted dental movement)
- integrate a wide choice of devices and anchorages (e.g., fixed, segmented techniques, miniscrew, headgear)
- compare different therapeutic options (e.g., with and without dental extractions, maxillofacial surgery)
- compare the treatment outcomes by superimposing the initial and final clinical situations (i.e., initial and final 3D digital patients dental arches).

V. Further thoughts

The superiority of a dynamic over a static presentation for learners' understanding and learning is debated in the literature. However, the animations and/or interactive medium could enhance the understanding of the effects that an orthodontic appliance has on the tooth movement. Animations are not sufficient, simulations' aims at fostering learning through immersion, reflection, feedback and practice minus the risks inherent to a real-life experience (i.e., to safer patient care) [16][6][12].

To assess the effectiveness of an innovative simulation device (in terms of understanding, gesture mastering and memorization), further studies on this subject should combine (i) an ergonomic approach, through a user-centered design, to identify the practitioners' needs and characteristics, (ii) an instructional engineering/educational psychology approach to design efficient learning activities.

References

[1] A. Mailloux, "Design of an innovative simulation device dedicated to the learning of biomechanics applied to orthodontics," *IARIA COGNITIVE, Nice*, June 2023.

[2] A. Mailloux., "Analysis of the needs to design an innovative device for distance continuing education for the use of practitioners in Orthopedics Dento-Facial : Contribution of a community of practice analysis," Doctoral thesis, Université de Lorraine, Dec. 2022.

[3] J. Faure, *"Orthodontic Biomechanics"*, sid ed., ser. The fundamentals, 2013.

[4] J.-M. Boucheix and J.-F. Rouet, "Are multimedia interactive animations effective for learning?" *French journal of pedagogy. Educational research*, no. 160, pp. 133–156, Sep. 2007.

[5] K. Sipiyaruk, P. Kaewsirirat, and P. Santiwong, "Technology-enhanced simulation-based learning in orthodontic education: A scoping review," *Dental Press Journal of Orthodontics*, vol. 28, p. e2321354, Jul. 2023.

[6] F. Ridzuan, G. K. L. Rao, R. M. A. Wahab, M. M. Dasor, and N. Mokhtar, "Enabling Virtual Learning for Biomechanics of Tooth Movement: A Modified Nominal Group Technique," *Dentistry Journal*, vol. 11, no. 2, p. 53, Feb. 2023.

[7] Y.-C. Lo, G.-A. Chen, Y.-C. Liu, Y.-H. Chen, J.-T. Hsu, and J.-H. Yu, "Prototype of Augmented Reality Technology for Orthodontic Bracket Positioning: An In Vivo Study," *Applied Sciences*, vol. 11, no. 5, pp. 1–13, Jan. 2021.

[8] A. Liebermann and K. Erdelt, "Virtual education: Dental morphologies in a virtual teaching environment," *J. Dent. Educ.*, vol. 84, pp. 1143–1150, 2020.

[9] E. Soltanimehr, E. Bahrampour, M. M. Imani, F. Rahimi, B. Almasi, and M. Moattari, "Effect of Virtual versus Traditional Education on Theoretical Knowledge and Reporting Skills of Dental Students in Radiographic Interpretation of Bony Lesions of the Jaw," *BMC Med. Educ*, vol. 19, no. 233, Jun. 2019.

[10] S. Mohammad Ali, J. Vakhnovetsky, and N. Nadershahi, "Scoping review of artificial intelligence and immersive digital tools in dental education," *J Dent Educ*, vol. 86, Jun. 2022.

[11] S. Schorn-Borgmann, C. Lippold, D. Wiechmann, and T. Stamm, "The effect of e-learning on the quality of orthodontic appliances," *Advances in Medical Education and Practice*, vol. 6, pp. 545–552, Aug. 2015.

[12] D. A. Cook, R. Hatala, R. Brydges, B. Zendejas, J. H. Szostek, A. T. Wang, P. J. Erwin, and S. J. Hamstra, "Technology-enhanced simulation for health professions education: a systematic review and meta-analysis," *JAMA*, vol. 306, no. 9, pp. 978–988, Sep. 2011.

[13] A. Wagner, W. Krach, K. Schicho, G. Undt, O. Ploder, and R. Ewers, "A 3-dimensional finite-element analysis investigating the biomechanical behavior of the mandible and plate osteosynthesis in cases of fractures of the condylar process," *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontics*, vol. 94, no. 6, pp. 678–686, Dec. 2002.

[14] D. Wagner, "Quantification and modeling of forces and moments applied inside orthodontic brackets placed on a dental arch in the three dimensions of space," Doctoral Thesis, Strasbourg, Jul. 2018. [Online]. Available: https://www.theses.fr/2018STRAD020

[15] P. Auconi, T. Gili, S. Capuani, M. Saccucci, G. Caldarelli, A. Polimeni, and G. Di Carlo, "The Validity of Machine Learning Procedures in Orthodontics: What Is Still Missing?" *Journal of Personalized Medicine*, vol. 12, no. 6, p. 957, Jun. 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9225071/

[16] M. Betrancourt, "Chapter 18. The Animation and Interactivity Principles in Multimedia Learning," *The Cambridge handbook of multimedia learning*, Jan. 2005.

Evaluating the Impact of Machine Learning Platforms on Cancer Classification Model Performance: A Cross-Platform Comparative Study

Adedayo Seun Olowolayemo
School of Engineering, Technology, and Design
Canterbury Christ Church University (CCCU)
Canterbury, UK
a.olowolayemo502@canterbury.ac.uk

Amina Souag
School of Engineering, Technology, and Design
Canterbury Christ Church University (CCCU)
Canterbury, UK
amina.souag@canterbury.ac.uk

Konstantinos Sirlantzis
School of Engineering, Technology, and Design
Canterbury Christ Church University (CCCU)
Canterbury, UK
Konstantinos.sirlantzis@canterbury.ac.uk

***Abstract* — Machine Learning techniques have become pivotal in advancing predictive models for early cancer detection, addressing the growing need for improved diagnostic efficiency. However, the role of implementation platforms in influencing model performance remains underexplored, even as variations in performance with the same dataset raise questions about platform choice. This study evaluates the impact of three ML implementation tools, the Scikit-learn, KNIME, and MATLAB on the performance of four classification algorithms: Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. Using the publicly available Wisconsin Diagnostic Breast Cancer dataset, these algorithms were implemented under default configurations and compared across key metrics: accuracy, recall, precision, and F1-score. Results revealed significant platform-dependent variations: Scikit-learn achieved consistently higher recall, particularly for Random Forest and Gradient Boosting, making it more effective at minimising false negatives critical in cancer diagnosis. MATLAB demonstrated superior precision, especially for Random Forest and Gradient Boosting, indicating potential in reducing false positives. KNIME, while effective in specific contexts, underperformed in recall and precision, raising concerns in scenarios requiring high sensitivity and specificity. These findings underscore the importance of platform selection based on predictive task requirements, especially in healthcare, where balancing false positives and false negatives is crucial. The study provides actionable insights for selecting ML platforms to enhance diagnostic accuracy in cancer classification tasks, with source code and data fully accessible through a public GitHub repository.**

Keywords - Cancer; Machine Learning; Python Scikit-learn; KNIME; MATLAB; Wisconsin Diagnostic Breast Cancer.

I. INTRODUCTION

Cancer remains a significant global health threat, causing nearly 10 million deaths in 2020 approximately one in six deaths globally underscoring its devastating impact and the urgent need for more effective prevention, early detection, and treatment strategies [1][2][3][4]. According to the World Health Organization (WHO), the disease affects individuals of all ages, including about 400,000 children each year. Notably, breast, lung, and colorectal cancers had the highest incidence rates, with lung cancer leading in mortality, followed by colorectal, liver, stomach, and breast cancers, as shown in Fig. 1. This figure depicts the distribution of new cancer cases and cancer-related deaths by type for 2020, highlighting the global burden of specific cancers and emphasizing the importance of early diagnosis and screening to reduce mortality and mitigate the far-reaching impacts of the disease [5].

Cancer arises from the uncontrolled division of cells, resulting in the formation of *tumours* that are classified as either *malignant* or *benign*. Malignant tumours are of particular concern due to their ability to invade surrounding tissues and spread to other parts of the body through metastasis [6]. This invasive behaviour complicates treatment, often requiring a combination of surgery, chemotherapy, and radiation [7]. Once metastasis occurs, malignant cells can establish secondary tumours in distant organs, such as the lungs, brain, or liver, further increasing the complexity of treatment and affecting patient prognosis [8].

In contrast, benign tumours, though also characterised by abnormal cell growth, remain localised and do not spread to other areas of the body. While generally less harmful, they can still pose risks depending on their size and location, particularly if they press on critical organs or tissues [9]. Treatment for benign tumours is typically less aggressive, though surgical removal may be necessary in cases where they cause discomfort or complications.

The distinction between malignant and benign tumours is crucial in understanding cancer progression, as well as the urgency and approach to treatment. This disease's complexity spans multiple organs including the breast,

kidneys, brain, lungs, prostate, ovaries, and skin, hence posing significant challenges for healthcare professionals. Despite advances in treatment, timely diagnosis remains critical; delays can lead to advanced stages of cancer that are more difficult to treat and are often associated with higher mortality rates.

Scientists are increasingly directing significant resources toward revolutionising the cancer diagnostic process, recognizing that early and accurate diagnosis can drastically improve patient outcomes. In this endeavour, Artificial Intelligence (AI) has emerged as a key player, demonstrating its potential across various domains, and now offering promising solutions in healthcare [10]. What sets AI apart in the medical field, particularly in cancer diagnosis, is its capacity to process and analyse vast amounts of complex data at speeds and scales that far exceed human capabilities.

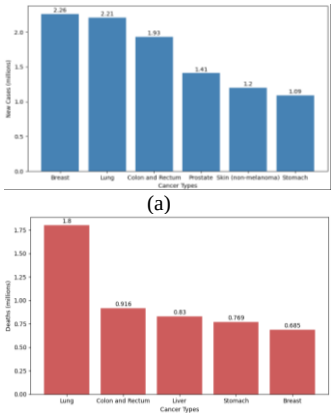


Figure 1. Chart of (a) New cancer cases by cancer type and (b) Cancer deaths by cancer type in 2020 [5]

Machine Learning (ML), a subset of AI, is not only enhancing efficiency but also transforming the nature of medical research. This transformation is evident in numerous studies [11][12][13], where AI techniques have been employed for the classification and prediction of cancer, as well as patient survival outcomes. Due to their ability to learn from data, ML algorithms trained on datasets can identify patterns and markers often imperceptible to human observers. [14]. This has led to breakthroughs in diagnostic precision, allowing for more accurate differentiation between diseases, including cancerous and non-cancerous conditions. Beyond diagnosis, ML is being leveraged to improve prognostic accuracy by predicting disease progression and response to treatment [15], helping health professionals make more informed decisions tailored to individual patients.

As ML continues to evolve, its impact extends beyond speed and precision; it has the potential to reshape the entire framework of cancer care. By integrating AI tools into clinical workflows, the hope is not only to expedite the diagnostic process but to also develop a more personalised, data-driven approach to treatment, where ML models help guide therapeutic choices with unprecedented accuracy. This shift represents a fundamental transformation in the healthcare industry, where the convergence of data science and medical practice could lead to faster, more reliable diagnoses and ultimately, improved survival rates for cancer patients.

By harnessing advanced computational techniques, ML algorithms ranging from Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), and Gradient Boosting (GBoost), among several others used for cancer diagnosis, extract insights from intricate medical data used in revolutionising clinical decision-making and improving patient outcomes from pinpointing diseases through image analysis [16] to forecasting patient responses to therapies.
However, a critical aspect that we found to be underexplored is the impact of implementation platforms on which the algorithms are trained, and models are developed, such as Python Scikit-learn, KNIME, and MATLAB on the performance of these algorithms. Therefore, understanding the nuanced influence of implementation platforms on ML algorithms is pivotal.

Against this backdrop, this study employed supervised learning to train models on Wisconsin Diagnostics Breast Cancer (WDBC) dataset [17] to evaluate the performance metrics of ML algorithms including accuracy, precision, recall, and F1-Score. The focus was on understanding the nuanced relationship between implementation platforms and the efficacy of these algorithms, emphasizing the potential impact of platform choice on algorithm behavior and highlighting the need to discern these disparities.
To achieve this, the study addresses two pivotal inquiries:

(1) It seeks to answer whether the choice of the implementation platform impacts the performance of ML algorithms in cancer data classification, and

(2) identifies which of the selected algorithms performed best in cancer dataset binary classification task.

This study delves into the complex interaction between ML algorithms, the platforms on which they are implemented, and the significance of the features within the dataset. Rather than focusing on optimising hyperparameters, the research aims to unearth deeper, more fundamental insights into how platform-specific factors influence model outcomes, including accuracy, efficiency, and predictive robustness. By utilizing the WDBC dataset, the study trains ML models to classify tumours as malignant or benign, a task critical for early cancer diagnosis. The focus on platform comparison allows for an exploration of how the underlying architecture and computational efficiency of different ML platforms can affect model performance, independent of tuning techniques.

This approach highlights a broader issue in ML research, how the choice of implementation tools can shape results beyond mere algorithm selection or dataset quality.
In analysing platform impacts on diagnostic accuracy, this study offers valuable contributions to developing more reliable and consistent ML-driven diagnostic systems, ensuring that performance improvements in cancer detection are not limited by the tools used to implement them. Ultimately, these findings provide a roadmap for more informed choices when developing ML models for medical applications, paving the way for advancements that can directly enhance patient outcomes.

The rest of this paper is organised as follows: Section II reviews related works, exploring the use of machine learning in cancer research. It examines studies that have applied ML algorithms, focusing on their implementation methods, train-test split strategies, performance evaluation metrics, dataset sources, and platforms utilised. Section III outlines the methodology adopted in this study, providing a detailed account of data collection, preprocessing, feature selection, and the implementation of selected ML models. Section IV presents the results, supported by an in-depth discussion of their implications. Finally, Section V concludes the paper by summarizing the findings and proposing directions for future research.

II. Related work

Researchers have explored and reported the use of various supervised ML algorithms in different areas of human health and medical fields. Some previous studies reviewed are briefly discussed below.

A. ML in Cancer Research

ML is reshaping the landscape of cancer research by offering powerful tools to improve key areas such as cancer classification and treatment outcome prediction. With its ability to process and analyse large amounts of data more efficiently, ML has allowed researchers to uncover patterns and insights that were previously out of reach, leading to more precise diagnoses and predictions. This section delves into a range of studies that demonstrate the practical benefits of ML in cancer research, shedding light on how it has enhanced diagnostic accuracy and predictive modeling. Despite these strides, there is still room to explore and fine-tune its applications. Through this review, we aim to highlight both the significant progress made and the opportunities for further development, emphasizing the potential for ML to drive even more impactful breakthroughs in cancer treatment and care.

Michael et al. in [18] tested five ML classification algorithms on 912 breast ultrasound images and found that Light Gradient Boosting Machine (LightGBM), the algorithm proposed in their work, which has an accuracy of 99.86%, outperformed other algorithms including K-Nearest Neighbour (KNN), and RF in binary classification of cancerous cells as either malignant or benign. Similarly, Ara et al. in [19] used a ML techniques to develop model for classifying cancer cells into two main categories. Kumar et al. in [20] on the other hand focused on using ML ensemble techniques for breast cancer detection and classification.
Their Optimized Stacking Ensemble Learning (OSEL) model showed a higher accuracy in performing the task than other ensemble ML techniques, such as Stochastic GBoost and XGBoost tested in their research. Ebrahim et al. [21] tested eight predictive algorithms on the National Cancer Institute dataset to identify which algorithm would predict cancer cell more accurately.

B. Selection of Algorithm

In cancer research involving ML, the selection of algorithms is a critical factor that can influence model performance, especially when applied to widely used datasets like the WDBC dataset. Numerous studies have utilised various ML algorithms for tasks such as classification, prediction, and diagnosis. This section reviews the algorithms commonly selected in existing literature, with a particular focus on those used in cancer research. While the current study aims to investigate how implementation tools may impact model performance under default settings, the literature at this stage primarily explores algorithm selection based on factors such as accuracy, ease of use, and compatibility with specific datasets. By examining these studies, we aim to uncover potential reasons behind the popularity of certain algorithms in the context of cancer classification, which can serve as a foundation for understanding the broader landscape of ML applications in healthcare.

LR, a linear model is a powerful predictive analysis tool that is especially useful for binary classification [22]. Zhu et al. in [22] experimented with improved LR in the classification of binary variable and independent variables to predict diabetes. Rahman et al. [23] examined six ML algorithms for predicting Chronic Liver Disease (CLD) and found the LR algorithm to be the most effective in predicting CLD based on the selected features.

Likewise, Tree based algorithms including DT, RF and GBoost are widely researched with the intent of harnessing their strengths particularly in performing classification tasks. Decision Trees (DT) provide a simple and interpretable framework upon which more advanced tree-based models, like Random Forests (RF) and Gradient Boosting (GBoost), are built, partitioning feature spaces into hierarchical branches to effectively capture non-linear relationships and feature interactions, enabling straightforward visualisation of decision-making processes. Moving beyond individual trees, RF combines multiple DTs through ensemble techniques, mitigating overfitting and increasing predictive accuracy [24]. By combining varied perspectives from individual trees, RF provides robust generalization and robustness to noisy data.

By extension, the GBoost algorithm, a more advanced method, embraces an iterative refinement to enhance predictive performance and in particular, Gradient Boosting Trees, such as XGBoost employ sequential tree fitting to target the residuals of prior iterations, systematically improving model predictions. These algorithms perform well in modeling complex relationships, accommodating non-linearities, and excelling in predictive accuracy across domains [25][26]. These characteristics formed the basis on which we selected the algorithms in our study.

C. Train-Test Split

The train-test split is a widely used method in ML, essential for assessing and comparing different algorithms or model configurations. By partitioning a dataset into two segments with one for training and one for testing, it ensures that models are evaluated consistently across the same testing subset. This process provides an unbiased framework for determining how well each model performs, free from the influence of the training data. Metrics such as accuracy, precision, recall, and F1-score, calculated from the test data, offer valuable insights into a model's potential performance in practical, real-world applications. More than just facilitating

model training, the train-test split underscores the necessity of rigorous validation to guarantee that the model's predictions are not only accurate but also reliable when deployed.

For evaluation, datasets used in various studies are split into different proportions using the larger proportion to train algorithms while the smaller proportion is used to test at the inference stage of model development. In [22], the authors assessed the performance of some classical and deep learning algorithms used to predict breast cancer, including DT, LR, KNN, Support Vector Machine (SVM), Recurrent Neural Networks (RNN) and Ensemble Learning. They used Train/Test split of 70:30 and 90:10. DT and Ensemble methods showed higher accuracy both before and after feature selection. Whereas DT did not perform optimally in predicting Kidney Cancer Lung Metastasis, as reported by [27], when trained with 52,222 records from the Surveillance, Epidemiology, and End Results (SEER) database and 492 hospital patient records with Train/Test split of 70:30 returning accuracy of 82% which is significantly lower than in other studies reviewed.

D. Performance Metrics

Efficient model development and deployment require a thorough evaluation to ensure reliable performance in real-world applications. One of the most essential tools in this process is the confusion matrix, which is a tabular representation that summarises the model's predictions against the actual outcomes, giving a detailed breakdown of a model's predictions [28]. It classifies outcomes into four categories: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) as we have illustrated in Fig. 2. By analysing these classifications, the confusion matrix helps reveal where a model excels and where it has not performed as expected or optimally. This level of detailed insight is particularly important in domains like healthcare, where incorrect predictions can have serious consequences, such as misdiagnoses or missed critical conditions.

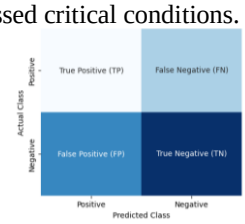


Figure 2. Illustrative Confusion Matrix Table.

Various important performance metrics are obtained from the confusion matrix, each providing a unique perspective for evaluating a model's effectiveness. The most commonly used metric is accuracy, which measures the proportion of correct predictions relative to the total number of predictions. While accuracy provides a broad overview of a model's success, it can be misleading, especially in scenarios with imbalanced datasets.

To address the limitations of accuracy, additional metrics such as precision, recall, and F1-score become crucial. Precision, which measures the proportion of correct positive predictions out of all positive predictions, is particularly relevant when the cost of false positives is high. In medical settings, a false positive incorrectly identifying a healthy individual as sick can lead to unnecessary treatments, anxiety, and strain on healthcare resources. Thus, high precision is essential to minimise the risk of falsely diagnosing healthy patients.

The performance metrics derived from the confusion matrix are computed based on equations (1-4) below.

$$Precision = TP / (TP + FP) \quad (1)$$

$$Recall = TP / (TP + FN) \quad (2)$$

$$F1\ Score = 2 * (Precision * Recall) / (Precision + Recall) \quad (3)$$

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (4)$$

Conversely, recall focuses on the model's ability to capture all actual positive cases, which is especially important in ensuring that no dangerous conditions are missed. In healthcare, missing a diagnosis, such as failing to detect cancer, can have devastating consequences. Therefore, high recall ensures that all true positive cases are identified, reducing the risk of underdiagnosis in critical conditions.

Balancing precision and recall is where the F1-score proves invaluable. The F1-score is the harmonic mean of precision and recall, offering a balanced evaluation of a model's ability to minimise both false positives and false negatives. This is particularly useful in datasets with class imbalances, where optimising for either precision or recall alone may not provide an accurate reflection of the model's true performance. In medical diagnostics, where both overdiagnosis (false positives) and underdiagnosis (false negatives) can have significant consequences, the F1-score serves as a comprehensive measure that helps to ensure models perform well across the spectrum of possible outcomes.

Ultimately, the confusion matrix and its associated metrics, (accuracy, precision, recall, and F1-score) offer a robust framework for assessing ML models, particularly in sensitive fields like healthcare. These metrics provide a deeper understanding of how models perform in various scenarios, ensuring they are not only accurate but also effective in minimising the risks associated with false predictions. By going beyond basic accuracy, this approach helps build trust in model deployment, ensuring that ML systems can reliably make critical decisions in complex, real-world environments.

Accuracy measures the proportion of correctly predicted instances in the dataset, providing a general overview of predictive success. Precision focuses on correctly predicted positive cases, which is crucial in scenarios like medical diagnoses where false positives can lead to adverse consequences. Recall assesses the model's ability to identify all true positive cases, essential for avoiding missed diagnoses in critical medical conditions. The F1-score balances precision and recall, offering a nuanced evaluation that is particularly useful for datasets with class imbalances. These four metrics collectively provide a comprehensive assessment of a model's performance.

E. Datasets

Data quality is fundamental in ML, shaping model development and real-world utility. The WDBC [17] has been pivotal in healthcare, especially for binary tumour classification, crucial in timely cancer detection and treatment planning. While a number of studies like [17][18][19] employed smaller, open-source WDBC datasets (typically fewer than 600 records and 30 features), other studies in [22] and [15] diverged. For example, [22] used a substantial dataset from the National Cancer Institute (NIH) containing 1.7 million records and 210 features. Despite its size, dataset quality, marked by precision and representativeness, significantly influences outcomes. Smaller datasets with these qualities outperform larger, noisier ones. This distinction is evident in accuracy rates, with open-source datasets achieving 99.12%, 99.67%, and 100%, compared to the model in [22] with a lower accuracy of 97.4%.

F. Implementation Platform

KNIME Analytics, a no-code tool known for its user-friendly interface and extensive integration with external tools, has been widely used in ML research, including studies such as [29], which explored cancer incidence among individuals with HIV in Zimbabwe. KNIME's appeal lies in its accessibility, allowing researchers without advanced programming skills to build and implement complex ML models. Meanwhile, Python, particularly with its rich ecosystem and powerful libraries like Scikit-learn, has established itself as a go-to platform for ML. Multiple studies, such as those in [30][31][32] have employed Python for cancer research, leveraging its versatility and the ability to fine-tune models through code.

In addition to KNIME and Python, MATLAB has also been widely used in ML research. Known for its robust computational capabilities, MATLAB offers a range of toolboxes and functions for developing ML algorithms. Its application in cancer diagnosis and classification tasks has been demonstrated across various studies, where it has been employed to build predictive models and evaluate performance across different classification algorithms. All three platforms including KNIME, Python (Scikit-learn), and MATLAB have significant backing from the scientific community. Each platform offers unique strengths, making it important to understand how platform-specific architectures and tools impact ML algorithm performance, especially in sensitive fields like cancer research.

The findings from the literature are summarised in Table I, which provides a comprehensive overview of recent studies utilizing ML techniques in cancer research. The table outlines critical aspects of each study, including the data sources, train-test split ratios, implementation platforms, algorithms employed, and resulting model accuracy. This summary enables a clear comparison of the approaches and outcomes in applying ML to cancer diagnosis and prognosis, offering insights into the varied impacts of different platforms and algorithms on model performance. (*a '–' has been used in the table to indicate instances where the relevant information was not available in the literature*).

TABLE I. COMPARATIVE REVIEW OF SOME STUDIES THAT USED ML TECHNIQUES IN CANCER RESEARCH.

Author, Year	Data Source	No of Records /Features	Train/Test Split	Implementation Platform	Algorithm Type	Model Accuracy
Ara et al. [19] , 2021	UCI	569/30	75:25	-	SVM, LR, KNN, DT, NB, RF	96.5%
Ebrahim et al. [21],2023	National Cancer Institute (NIH),	70,079/107	70:30 &90:10	- Python	DT, LR, VM, LD, ET, KNN	98.7%
Minnoor et al.[24] 2023	UCI	569/30	80:20	-	RF, SVM, DT, MLP, KNN	100%
Yi et al., [27],2023	SEER& Southwest Hospital, China.	52,714 / -	70:30	Python	LR, XGBoost, RF, SVM, ANN, DT	-
Shafique et al.[29],2023	Kaggle	569/30	75:25		RF, VM, GBoost, LR, MLP, KNN	100%
Uddin et al. [30], 2023	UCI	569/30	70:30	Python	SVM, RF, KNN, NB, DT, LR, AB, GBoost, MLP, NCC, VC	98.7 %
Mahesh et al., [33],2022	Kaggle	143/10	70:30	Python	NB, AltDT, RedEPT, RF	98.20%
Zhang et al [34]., 2022	TCGA	604/ -	-	R & Python	RF, SVM, libD3C	99.67%
Aamir et.al.[35], 2022	UCI	569/26	80:20 &70:30	Python & Tensor Flow	RF, GBoost, SVM, ANN, MLP	99.12%
ATEŞ et al. [36] 2021	Kaggle	569/30	70:30	KNIME	NB, DT, MLP	96.5%
Liu, et al. [37]2018	UCI	569/30	75:25	Python	LR	96.5%
Dora et al., 2017 [45]	UCI	569/30	70:30	MATLAB	ELM	94.52%

While numerous studies have demonstrated the effectiveness of machine learning (ML) in cancer diagnosis and classification, a critical gap persists in understanding how the choice of implementation platform influences model performance. Most research has focused on algorithm selection and dataset quality, operating under the assumption of platform independence. This neglects potential disparities introduced by differences in platform architectures, default configurations, and computational efficiencies, which could significantly affect model outcomes and their broader applicability. Addressing this unexplored area forms the crux of our research.

By systematically investigating how different ML implementation platforms including KNIME, Scikit-learn, and MATLAB impact the performance of widely used classification algorithms in cancer diagnostics, we aim to shed light on a critical yet overlooked factor. Our study explores platform-dependent variations across key metrics such as accuracy, recall, precision, and F1-score, offering a novel perspective that underscores the importance of platform choice in high-stakes applications like healthcare.

The practical implications of this research are substantial as understanding how platform-specific characteristics influence model accuracy, efficiency, and scalability enables

more informed decisions in real-world applications where performance optimisation is essential.

By providing actionable insights and a rigorous methodological framework, this study contributes to the broader discourse in ML research, encouraging further consideration of technical environments and fostering advancements that improve diagnostic workflows and patient outcomes. Ultimately, this work fills a vital gap in existing literature and establishes a foundation for optimising ML workflows across diverse computational platforms

In the sections that follow, we delve into the methodology designed to rigorously test this hypothesis, offering a framework that enables a deeper understanding of how platform-specific characteristics may impact model performance across various algorithms. This novel perspective enhances the current discourse in ML, encouraging further consideration of the technical environments in which models are deployed.

III. METHODOLOGY

This study's methodology comprises systematic steps for a comparative analysis of ML algorithms using the WDBC dataset and three implementation platforms. The process as illustrated in Fig. 7 includes data collection, exploration, feature engineering and selection using filtering and RF techniques. The dataset was split into an 80% training set and a 20% test set before model development, ensuring a robust evaluation process.

A. Data Collection and Preprocessing

We selected the publicly available WDBC dataset from the University of California, Irvine (UCI) ML repository [18] because of its origin in medical research, extensive use in breast cancer-related ML studies, and established reputation in the research community. Its real-world applicability makes it a reliable choice for binary classification tasks. The dataset contains 569 instances and 30 attributes, extracted from digitised Breast Mass Fine Needle Aspiration (FNA) specimens. These attributes include measurements such as "radius_mean," "texture_mean," and "perimeter_mean," which represent features of cell nuclei in biopsy images.

The dataset is divided into two classes: benign tumours, comprising 62.7% of the total instances, and malignant tumours, making up the remaining 37.3%. We show in Fig. 3 the proportion of these two classes, highlighting the distribution of benign and malignant cases for further analysis [18].

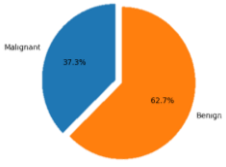


Figure 3. Pie chart showing percentage composition of the class labels as M-malignant and B-benign.

Following the dataset analysis, we conducted a correlation analysis to explore relationships between features, as illustrated in the heatmap in Fig. 4. This step is crucial for feature selection, offering insights into how each feature correlates with the target variable and other features. Identifying multicollinearity when features are highly correlated is essential, as redundant features can complicate the model without enhancing predictive performance. This process helps ensure the model remains efficient and effective.

The correlation analysis serves two purposes: identifying features strongly correlated with the target variable for their predictive potential and detecting pairs of highly correlated features. When features exhibit high correlation (close to ±1), removing one of them helps reduce redundancy and streamline the model without affecting performance.

In this study, where the aim is to investigate the impact of ML implementation platforms on model performance, optimising the feature set before comparing models is crucial. Since models are tested using default platform settings, including only the most relevant features becomes even more important. Retaining irrelevant or redundant features could obscure performance differences between platforms by introducing noise or inflating the models unnecessarily.

The correlation analysis assessed the relationships between features, providing insights into their relevance to the target variable and identifying interdependencies between them. The heatmap in Fig. 4 highlights these correlations, with darker shades indicating stronger relationships and lighter shades indicating weaker ones. This visual helps identify redundant features due to high correlation, guiding better feature selection decisions. Addressing such correlations improves predictive accuracy and reduces the risk of overfitting by ensuring a streamlined feature set. This process enhances the model's overall efficiency and reliability by retaining only the most relevant and independent features.

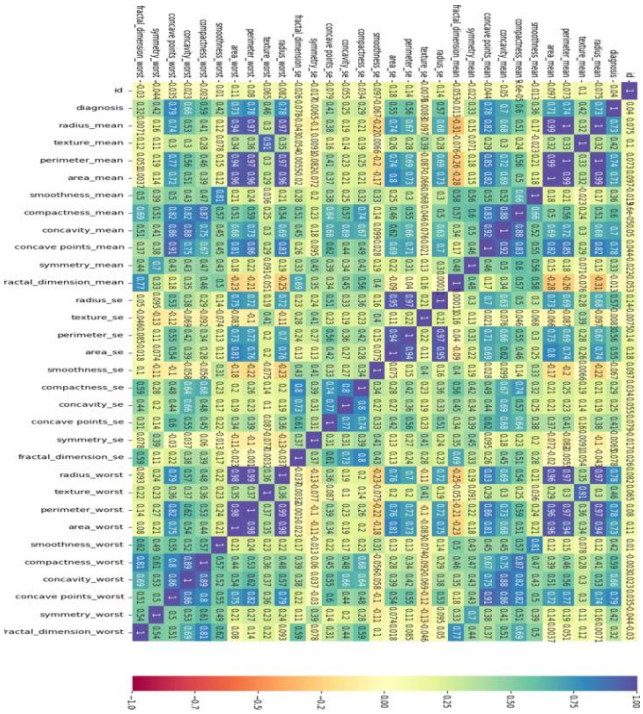


Figure 4. Heatmap of features correlation

In the data preprocessing phase, the dataset was structured into a Python dataframe which we subsequently queried to ascertain the data types and to check for presence of any null or missing values [38]. Table II extracted from our code implementation for Exploratory Data Analysis (EDA), confirms that the WDBC data contains a mix of integer and floating-point values, with no null values identified. Further analysis included detecting outliers using box plots, and the capping method was applied to mitigate their impact, ensuring the dataset's integrity for subsequent analyses. This technique, as presented by [39] involved setting values below the lower whisker to the lower whisker's value and values above the upper whisker to the upper whisker's value, ensuring an unbiased dataset.

TABLE II. WDBC DATASET VARIABLES DATATYPE.

```
Data columns (total 32 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       569 non-null    int64
 1   diagnosis                569 non-null    int32
 2   radius_mean              569 non-null    float64
 3   texture_mean             569 non-null    float64
 4   perimeter_mean           569 non-null    float64
 5   area_mean                569 non-null    float64
 6   smoothness_mean          569 non-null    float64
 7   compactness_mean         569 non-null    float64
 8   concavity_mean           569 non-null    float64
 9   concave points_mean      569 non-null    float64
 10  symmetry_mean            569 non-null    float64
 11  fractal_dimension_mean   569 non-null    float64
 12  radius_se                569 non-null    float64
 13  texture_se               569 non-null    float64
 14  perimeter_se             569 non-null    float64
 15  area_se                  569 non-null    float64
 16  smoothness_se            569 non-null    float64
 17  compactness_se           569 non-null    float64
 18  concavity_se             569 non-null    float64
 19  concave points_se        569 non-null    float64
 20  symmetry_se              569 non-null    float64
 21  fractal_dimension_se     569 non-null    float64
 22  radius_worst             569 non-null    float64
 23  texture_worst            569 non-null    float64
 24  perimeter_worst          569 non-null    float64
 25  area_worst               569 non-null    float64
 26  smoothness_worst         569 non-null    float64
 27  compactness_worst        569 non-null    float64
 28  concavity_worst          569 non-null    float64
 29  concave points_worst     569 non-null    float64
 30  symmetry_worst           569 non-null    float64
 31  fractal_dimension_worst  569 non-null    float64
dtypes: float64(30), int32(1), int64(1)
```

Normalization was achieved through Z-Score Normalization (Standardization). This rescales each feature to a normal distribution with a mean of 0 and a standard deviation of 1 [40][41]. Standardizing features to a common scale is a crucial step in ML to ensure that algorithms do not disproportionately favor features with larger magnitudes. This is especially important for gradient-based models like LR where unscaled features can skew the learning process. By applying z-score standardization (as shown in Equation 5), we normalised each feature to have a mean of zero and a standard deviation of one. This not only enhances the model's ability to learn balanced patterns but also improves the convergence speed during training. This step ultimately helps in improving fairness, accuracy, and overall model performance across a variety of ML algorithms [41].

$$Z=(x-\mu)/\sigma. \quad (5)$$

where z is the scaled value of the feature,
x is the original value of the feature,
μ is the mean value of the feature, and
σ is the standard deviation of the feature.

B. *Feature Selection*

In ML studies, selecting the most informative features is a critical step in optimising model performance. Among the many techniques available for feature selection, Spearman's rank correlation is a popular choice for identifying relevant features based on their effectiveness in handling datasets where relationships between variables and the target are not strictly linear, making it valuable in a wide range of ML tasks. By ranking features according to their correlation with the target, Spearman correlation helps filter out less important features, ultimately improving the model's accuracy and efficiency.

In this study, we implemented a two-step feature selection process utilizing both the Filter Method and the Tree-Based Method. Initially, the Filter Method applied Spearman rank correlation to evaluate the features based on their correlation coefficients with the target variable. Features with correlation coefficients ≤ 0.5 were deemed insignificant and removed, following the guidelines established in previous works by [44]. This threshold-based approach resulted in the selection of 15 out of the 30 original features, which were considered sufficiently relevant for further analysis.

Spearman's rank correlation, being a non-parametric measure, was used here because it can handle monotonic relationships without assuming a linear connection between features and the target variable.

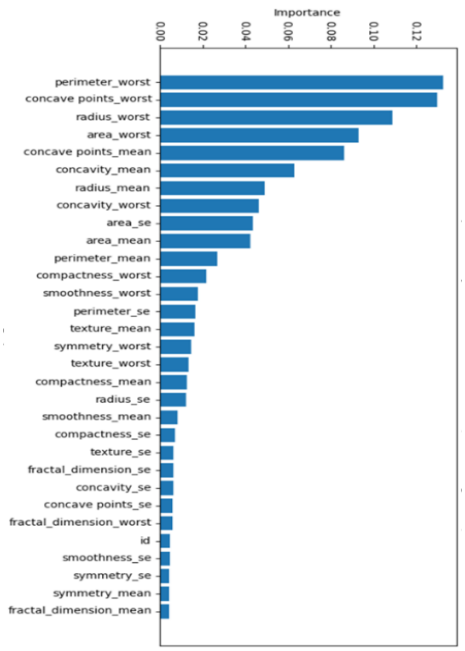


Figure 5. Random Forest features importance ranking, showing their importance.

Following the Spearman rank correlation and initial feature selection, we further validated the importance of the features using a RF classifier. The RF algorithm provided feature importance scores, highlighting the most influential variables for model development, as illustrated in Fig. 5 below. The top features were primarily geometric properties

of the tumour, such as perimeter_worst, concave points_worst, and radius_worst. These features, representing worst-case tumour measurements, played a critical role in distinguishing between classes, suggesting that extreme tumour characteristics are essential for accurate predictions.

In contrast, features such as fractal_dimension_mean, symmetry_mean, and smoothness_se were among the least important, contributing minimally to model performance. These features likely provided less useful information for classification, reaffirming the need to focus on features with higher predictive value.

This two-step approach involving the combination of spearman rank correlation with a tree-based method allowed us to filter out less relevant features while retaining those most critical for improving the model's predictive power. The results emphasise the importance of selecting features that capture key biological characteristics, particularly in contexts like cancer classification, where geometric properties of tumours are pivotal in distinguishing malignant from benign cases.

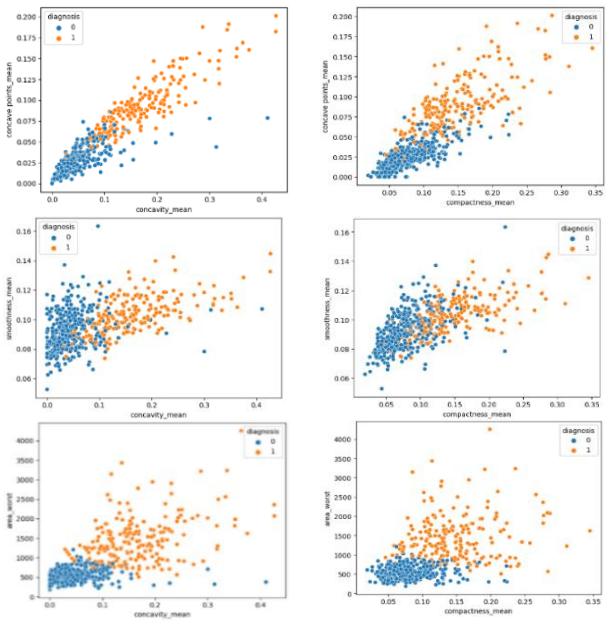


Figure 6. Scatter plot showing relationships between selected features. (Additional views of relationships between other features can be accessed in the GitHub repository [43]).

This method, known for balancing interpretability and computational efficiency while capturing both linear and non-linear relationships, affirmed the chosen features, as shown in Fig. 6, underscoring their significance in model development [42]. The synergy between the two methods ensured a comprehensive and accurate feature selection process, crucial for enhancing the model's predictive capabilities.

Understanding the relationship between the features helped to inform the class of ML algorithms that will be best suited for the classification task.

C. Model Selection and Implementation

In this study, four supervised ML classification algorithms were selected based on their unique attributes and widespread usage in previous research. LR was chosen for its ability to estimate outcome probabilities, making it a suitable choice for binary classification tasks. Its interpretability and computational efficiency further contribute to its popularity, as it provides a balance between performance and simplicity. On the other hand, DT, RF, and GBoost were selected for their ability to partition the data recursively. This recursive approach enables these algorithms to efficiently identify the most relevant features and optimal split points, which is crucial for improving classification accuracy.

The study was conducted using three platforms: KNIME Analytics Platform (Version 4.7.6), Python (Version 3.11.4, JupyterLab) with the Scikit-learn library, and MATLAB R2024a. For each platform, the ML algorithms were trained and tested using default settings, without any parameter tuning.

In KNIME, an exception was made for the RF algorithm, where the default split criterion was modified from "Information Gain Ratio" to "Gini Index." This adjustment was made to align with the default settings used in Scikit-learn, ensuring consistency and fairness in the comparative analysis. No such adjustments were made in MATLAB, as the platform's default configurations were retained for all algorithms. This approach allowed for a standardised comparison of the platforms, providing insights into how each platform handles the same ML models under comparable conditions.

To assess the algorithms' performance, the dataset was divided using an 80:20 train-test split. This split allocated 80% of the data for training, allowing the models to learn from the underlying patterns in the data, while the remaining 20% was used to test their ability to generalise to new, unseen instances. This approach provided a robust framework for evaluating the algorithms' effectiveness in classification tasks.

The source code and data used in this study, are available in a public GitHub repository to facilitate transparency and reproducibility. The methodology employed was designed to allow for a comprehensive evaluation of the selected algorithms while ensuring consistency in the comparative analysis [43].

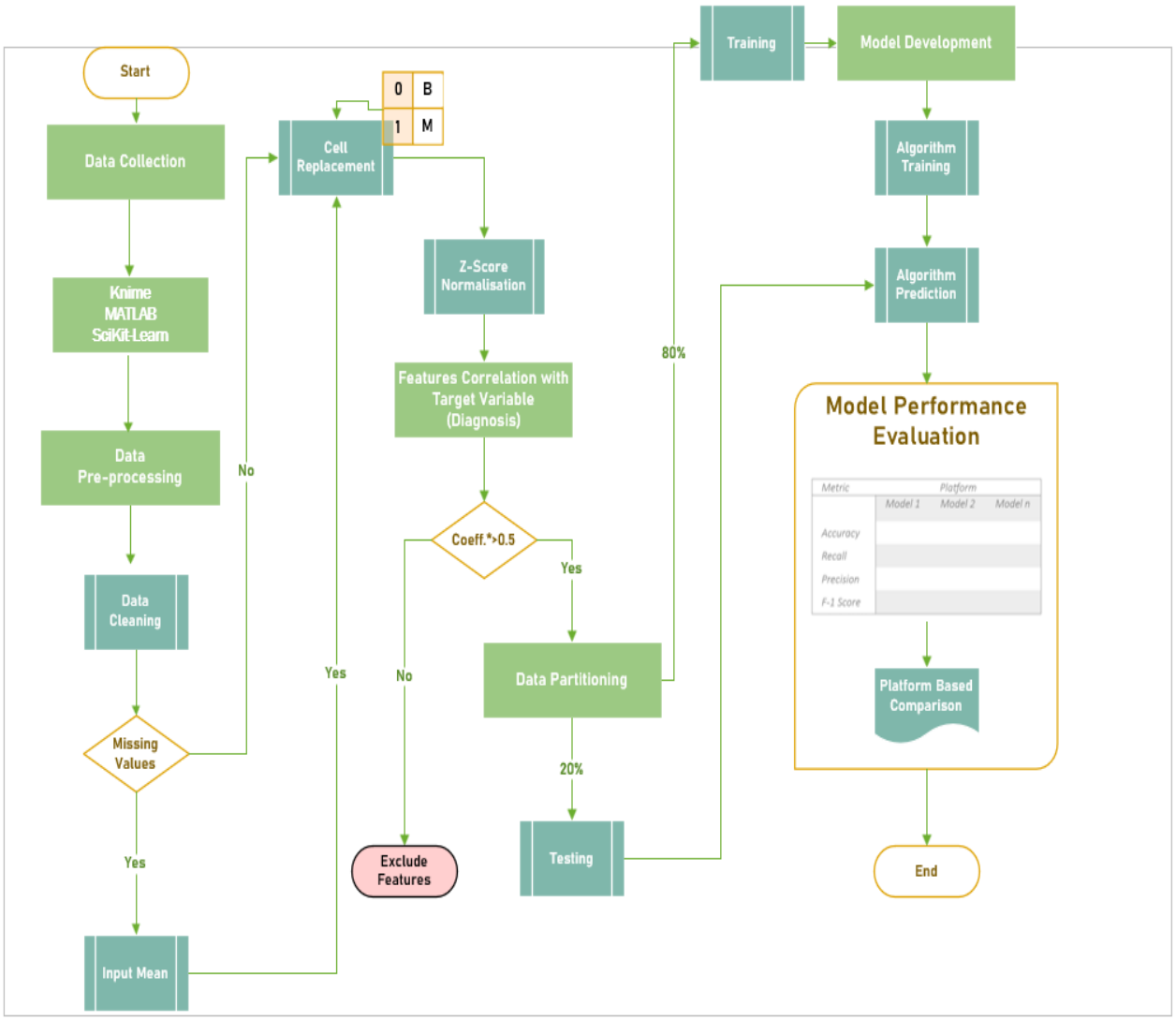


Figure 7. Flowchart illustrating the research methodology employed in this study.

IV. RESULTS AND DISCUSSION

This section outlines the experimental results obtained from implementing the four ML algorithms, LR, DT, RF, and GBoost across three platforms: Scikit-learn, KNIME, and MATLAB. The results are summarised in Table III and visually illustrated in Fig. 8, which depicts how these algorithms performed on the different platforms based on key metrics: Accuracy, Recall, Precision, and F1-Score. These metrics were used to evaluate and analyse the effectiveness of each algorithm in handling classification tasks across the platforms. The implementation of each platform was carefully examined to provide insights into how underlying differences in architecture and execution influence model performance.

A. Results Overview

Beginning with LR, Scikit-learn exhibited the highest overall performance across all metrics. An accuracy of 95.6%, combined with a recall of 0.929, precision being 0.951, and an F1-Score of 0.940, reflects the platform's ability to balance sensitivity and specificity under default settings. The high recall indicates that Scikit-learn's implementation is particularly effective at identifying true positives which is an important characteristic in healthcare scenarios where the misclassification of a malignant tumour as benign could delay treatment. Moreover, the precision score suggests that the platform manages to minimise false positives, which helps avoid unnecessary treatment for benign cases. Given that LR is a foundational algorithm, these results may reflect a strong alignment between the algorithm's mathematical structure and the default handling by Scikit-learn.

TABLE III. COMPARATIVE ASSESSMENT OF MODEL PERFORMANCE ON THE TWO PLATFORMS.

Algorithm	Tool	Accuracy	Recall	Precision	F1-Score
LR	Scikit-learn	0.956	0.929	0.951	0.940
	KNIME	0.921	0.884	0.905	0.894
	Matlab	0.938	0.921	0.897	0.909
DT	Scikit-learn	0.930	0.952	0.870	0.909
	KNIME	0.886	0.907	0.813	0.857
	Matlab	0.903	0.833	0.897	0.864
RF	Scikit-learn	0.947	0.976	0.891	0.932
	KNIME	0.912	0.884	0.884	0.884
	Matlab	0.956	0.925	0.949	0.937
GBoost	Scikit-learn	0.974	0.976	0.953	0.965
	KNIME	0.904	0.861	0.881	0.871
	Matlab	0.965	0.949	0.949	0.949

KNIME's performance for LR was comparatively lower, with an accuracy of 92.1%, recall of 0.884, precision of 0.905, and an F1-Score of 0.894. The lower recall indicates a reduced sensitivity to identifying positive cases, implying a higher rate of missed malignancies, which could have severe consequences in diagnostic applications. The precision, while reasonable, suggests that KNIME's default implementation may produce more false positives than Scikit-learn. This difference in the balance of sensitivity and specificity between platforms could have obvious practical implications in fields where both false negatives and false positives carry significant costs.

MATLAB's implementation of LR on the other hand showed an intermediate performance between the two platforms, with an accuracy of 93.8%, recall of 0.921, precision of 0.897, and an F1-Score of 0.909. Although MATLAB showed better recall than KNIME, indicating improved detection of true positives, its precision was lower than that of Scikit-learn suggesting that MATLAB's LR model may generate a higher number of false positives under default conditions, potentially leading to overdiagnosis in clinical settings. Despite this, the relatively balanced performance across all metrics indicates that MATLAB can still handle classification tasks effectively, albeit with slight trade-offs in sensitivity versus specificity.

The results from the DT algorithm reveal more noticeable disparities between platforms. Scikit-learn achieved an accuracy of 93.0%, recall of 0.952, precision of 0.870, and an F1-Score of 0.909, indicating a robust performance in classifying positive cases. The high recall suggests that Scikit-learn's DT model was able to identify most malignant cases, which is critical in ensuring that no critical diagnoses are overlooked. However, the slightly lower precision score points to a higher rate of false positives, meaning that some benign cases were misclassified as malignant, potentially leading to unnecessary medical interventions.

In contrast, KNIME demonstrated lower overall performance with DT, recording an accuracy of 88.6%, recall of 0.907, precision of 0.813, and an F1-Score of 0.857. The reduction in both precision and recall indicates that KNIME's DT model may struggle more with distinguishing between positive and negative cases under default settings. A lower precision suggests that false positives are more frequent, while the lower recall implies that true positives are being missed. This is particularly concerning in healthcare applications, where both types of errors can have significant consequences for patient care.

MATLAB's DT implementation performed with an accuracy of 90.3%, recall of 0.833, precision of 0.897, and an F1-Score of 0.864. While MATLAB's precision was higher than that of Scikit-learn, indicating fewer false positives, its lower recall shows that it missed more positive cases. This balance suggests that MATLAB's DT model, under default settings, may be more conservative, favoring the reduction of false positives but potentially at the expense of missing some malignant cases. In contexts where it is critical to detect as many positive cases as possible, this trade-off in favor of precision could impact decision-making.

Moving to the RF algorithm, both Scikit-learn and MATLAB exhibited strong performances with accuracy levels of 95.6%. However, Scikit-learn's recall (0.976) was notably higher than MATLAB's (0.925), suggesting that Scikit-learn's implementation was more sensitive to identifying true positives. This higher recall is particularly important in healthcare applications where failing to detect a malignant case could have serious consequences. In contrast, MATLAB's precision (0.949) exceeded that of Scikit-learn

(0.891), implying that MATLAB's RF model produced fewer false positives. This suggests that MATLAB's default RF implementation may prioritise specificity over sensitivity, which could be advantageous in cases where reducing unnecessary medical procedures is critical. The F1-Scores of 0.932 for Scikit-learn and 0.937 for MATLAB reflect a strong overall balance in their respective RF models.

KNIME's RF performance was lower, with an accuracy of 91.2%, recall of 0.884, precision of 0.884, and an F1-Score of 0.884. The equal precision and recall scores suggest that KNIME's RF model maintained a balance between sensitivity and specificity, but both were lower compared to Scikit-learn and MATLAB. The lower recall indicates that KNIME's RF model, under default settings, may miss more malignant cases, while the lower precision points to a higher rate of false positives. This trade-off could be significant in clinical applications where minimising both false positives and false negatives is essential.

TABLE IV. PLATFORM BASED CONFUSION MATRIX OF THE ALGORITHMS.

Scikit-learn

	LR		DT	
	Positive	Negative	Positive	Negative
Positive	70	2	66	6
Negative	3	39	2	40

	RF		GBoost	
	Positive	Negative	Positive	Negative
Positive	67	5	70	2
Negative	1	41	1	44

KNIME

	LR		DT	
	Positive	Negative	Positive	Negative
Positive	67	4	62	67
Negative	5	38	4	5

	RF		GBoost	
	Positive	Negative	Positive	Negative
Positive	66	5	66	5
Negative	5	38	6	37

MATLAB

	LR		DT	
	Positive	Negative	Positive	Negative
Positive	71	3	67	7
Negative	4	35	4	35

	RF		GBoost	
	Positive	Negative	Positive	Negative
Positive	71	3	72	2
Negative	3	36	2	37

The GBoost algorithm exhibited the largest performance differences across platforms. Scikit-learn achieved an accuracy of 97.4%, recall of 0.976, precision of 0.953, and an F1-Score of 0.965. These results suggest that Scikit-learn's default GBoost model is highly sensitive and effective at minimising both false positives and false negatives. High recall ensures that most positive cases are correctly identified, while high precision reduces the number of benign cases misclassified as malignant. This balance makes Scikit-learn's GBoost implementation suitable for applications where both sensitivity and specificity are crucial.

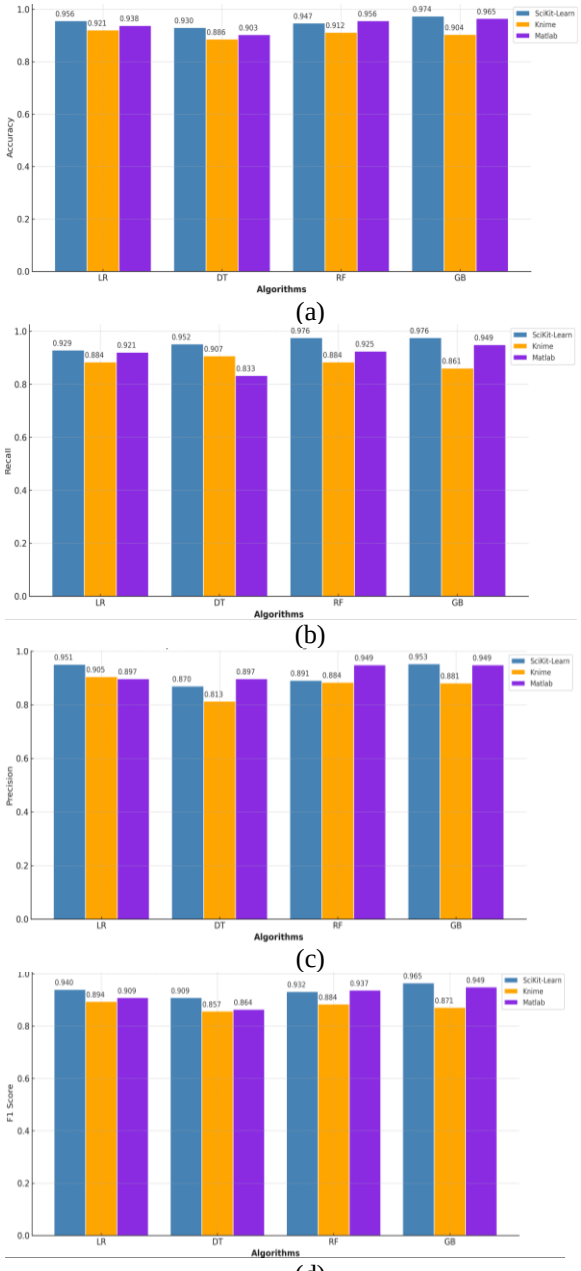


Figure 8. Column chart comparing the performance of all algorithms on the platforms for:
(a) Accuracy, (b) Recall, (c) Precision, and (d) F1-Score.

MATLAB also performed well with GBoost, achieving an accuracy of 96.5%, recall of 0.949, precision of 0.949, and an F1-Score of 0.949. While the results are close to those of Scikit-learn, MATLAB's slightly lower recall suggests that it may miss more positive cases, which could be critical in high-stakes applications like cancer detection. However, its

matching precision indicates that MATLAB is effective at reducing false positives, contributing to its balanced overall performance.

In comparison, KNIME's GBoost model showed lower performance, with an accuracy of 90.4%, recall of 0.861, precision of 0.881, and an F1-Score of 0.871. The lower recall value indicates that KNIME's model may miss more true positives, and the reduced precision suggests a higher rate of false positives compared to Scikit-learn and MATLAB.
This could impact diagnostic accuracy, especially in scenarios where identifying every possible positive case is crucial to patient outcomes.

Further insights into these performance metrics are provided by analysing the confusion matrices. Scikit-learn consistently demonstrated lower false negative (FN) rates compared to KNIME, particularly for RF and GBoost. For example, Scikit-learn's RF and GBoost models reported only 1 false negative each, while KNIME misclassified 5-6 malignant cases as benign. In the context of medical diagnostics, such discrepancies are significant, as false negatives can delay necessary treatments and worsen patient prognosis. Scikit-learn's lower false positive (FP) rates across all algorithms also suggest fewer benign cases misclassified as malignant, reducing the likelihood of unnecessary medical interventions and related costs. This trend was especially evident in the DT and RF models, where KNIME displayed higher FP rates, indicating that platform-specific characteristics might influence error rates in default implementations.

The study reveals performance variations in the algorithms tested across the platforms when executed with default settings. Scikit-learn consistently showed higher recall across all algorithms, particularly in RF and GBoost, where minimising false negatives is critical. MATLAB performed comparably in many instances but generally exhibited slightly lower recall, potentially missing more true positives. KNIME, while maintaining a balance between precision and recall, generally demonstrated lower performance in both metrics, especially in GBoost. These findings underscore the importance of considering the implementation platform when developing ML models, as platform-specific characteristics can influence how models handle classification tasks and the balance between sensitivity and specificity.

In the context of cancer care and ML research, analysing the confusion matrix presented in Table IV, which includes values for true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), provides critical insights into model effectiveness and potential clinical implications. For cancer diagnosis, the priority is to minimise false negatives (FN), as these represent cases where malignant tumours are misclassified as benign. In the presented tables, models implemented in Scikit-learn consistently have lower FN rates compared to KNIME, particularly with RF and GBoost. Scikit-learn's RF and GBoost models show only 1 false negative, whereas KNIME's implementations misclassified 5-6 malignant tumours as benign, raising concerns about its reliability in this critical area.

Equally important is the rate of false positives (FP), where benign tumours are mistakenly classified as malignant. While false positives are less harmful than false negatives, they still pose risks in cancer care by leading to unnecessary treatments, patient anxiety, and potential overtreatment. In this regard, Scikit-learn once again shows better performance, with fewer false positives across models compared to KNIME. For instance, Scikit-learn's LR and GBoost models have just 2 false positives each, whereas KNIME's counterparts show a higher rate, with up to 9 false positives in the DT model.

The true positives (TP) and true negatives (TN) in both platforms indicate the number of correctly classified malignant and benign cases, respectively. High TP values are essential in ensuring that patients with cancer receive timely treatment, while high TN values prevent unnecessary interventions for healthy patients. Scikit-learn demonstrates higher TP and TN rates overall, especially in the RF and GBoost models, where the identification of both malignant and benign cases is nearly flawless. This performance consistency highlights the critical importance of model accuracy and optimisation in ML research for cancer care, where minimising both FN and FP is essential for improving clinical outcomes.

B. Statistical Analysis

The results of this study reveal that the performance of classification algorithms can vary across machine learning platforms, even with consistent datasets and preprocessing steps. To validate these observations, the normality of performance metrics including Accuracy, Recall, Precision, and F1-Score was assessed using the Shapiro-Wilk test. All metrics satisfied the normality assumption, with p-values exceeding 0.05, enabling the use of parametric tests. Repeated measures ANOVA in Table V identified statistically significant differences in Accuracy ($p = 0.0054$) and F1-Score ($p = 0.0107$), suggesting that variations in these metrics are unlikely to be random and may be influenced by platform-specific factors. However, Recall ($p = 0.0730$) and Precision ($p = 0.0757$) did not show significant differences, indicating limited platform-specific effects on these metrics.

TABLE V. ANOVA TEST RESULTS

Metric	F Value	p-value	Significance
Accuracy	14.1037	0.0054	Yes
Recall	4.1794	0.073	No
Precision	4.0927	0.0757	No
F1-Score	10.6302	0.0107	Yes

TABLE VI. FRIEDMAN'S TEST RESULTS

Metric	Friedman Statistic	p-value	Significance
Accuracy	6.5	0.0388	Yes
Recall	6.5	0.0388	Yes
Precision	3.5	0.1738	No
F1-Score	6.5	0.0388	Yes

To further confirm these findings, Friedman's test, a non-parametric alternative, reinforced the ANOVA results by identifying significant variances in Accuracy, Recall, and F1-Score ($p = 0.0388$ for all), while Precision remained statistically insignificant ($p = 0.1738$) as seen in Table VI. Pairwise comparisons using Tukey's HSD test, shown in Table VII indicate that KNIME significantly underperformed compared to Scikit-learn in Accuracy, Recall, and F1-Score ($p = 0.0302$, 0.0311, and 0.0299, respectively). MATLAB exhibited no significant differences when compared to either platform, indicating comparable performance but neither superiority nor inferiority. Precision showed no significant differences across any platform pairs, highlighting its insensitivity to platform-specific factors in this context.

These results underscore that platform-specific characteristics, such as optimisation techniques and library implementations, play a significant role in influencing Accuracy and F1-Score but have minimal impact on Recall and Precision. Scikit-learn's superior Recall and F1-Score, particularly when compared to KNIME, highlight the importance of selecting platforms that consistently demonstrate high sensitivity and specificity in critical applications like healthcare. This study emphasises the importance of platform selection in machine learning research and applications, particularly in high-stakes domains like healthcare. It also sets the stage for further investigations into architectural and algorithmic factors driving these platform-dependent performance differences.

V. CONCLUSION AND FUTURE WORK

This comparative experiment examined the impact of different machine learning (ML) implementation platforms on the performance of classification models, focusing on four commonly used algorithms LR, DT, RF, and Gradient GBoost applied to the WDBC dataset. The analysis involved three platforms: Scikit-learn, KNIME Analytics, and MATLAB, and explored the behavior of these models under default configurations. Significant variations were observed across platforms, with each platform demonstrating unique strengths based on the metrics of accuracy, recall, precision, and F1-Score.

The study's findings highlighted that Scikit-learn consistently achieved high recall across algorithms like DT, RF, and GBoost, which is particularly important in healthcare applications such as cancer diagnosis, where minimising false negatives is crucial. The ability to correctly identify true positive cases ensures that malignant tumors are not overlooked, which is essential for timely treatment. In contrast, KNIME showed strong performance with LR, demonstrating higher accuracy for that algorithm but generally lower recall across other algorithms. This suggests that while KNIME may be effective in specific scenarios, its capacity to handle high-sensitivity tasks such as cancer detection, where recall is critical may be limited under default conditions.

MATLAB, meanwhile, presented a balanced approach, particularly excelling in precision with models like RF and GBoost, suggesting that it may be more suitable for applications where reducing false positives is important, such as in scenarios aiming to minimise unnecessary treatments. However, its lower recall compared to Scikit-learn suggests that it may miss more true positive cases, which could be a concern in healthcare settings which could delay treatment if a positive diagnosis is missed. These results emphasise the importance of selecting the right platform based on the specific objectives of a given ML task. The trade-offs between sensitivity (recall) and specificity (precision) can vary significantly depending on the platform, as demonstrated by the variations in performance across Scikit-learn, KNIME, and MATLAB. For applications such as cancer diagnosis, where both false positives and false negatives carry serious implications, platform choice is not just a technical consideration but a decision that can significantly influence model outcomes and, by extension, patient care.

The statistical analysis further underscores the significance of these platform-specific variations. Using the Shapiro-Wilk test, the normality of the performance metrics was confirmed, allowing parametric tests like repeated measures ANOVA to validate the significance of the observed differences. ANOVA identified statistically significant differences in Accuracy ($p = 0.0054$) and F1-Score ($p = 0.0107$), indicating that the variations in these metrics are unlikely to be random. Conversely, Recall ($p = 0.0730$) and Precision ($p = 0.0757$) did not exhibit significant differences, suggesting that platform-specific factors have minimal influence on these metrics.

These results were further validated using Friedman's test, which supported the significance of variations in Accuracy, Recall, and F1-Score while confirming that Precision remained statistically insignificant. Pairwise comparisons using Tukey's HSD test highlighted that KNIME significantly underperformed compared to Scikit-learn in Accuracy, Recall, and F1-Score, while MATLAB showed no significant differences across any metrics when compared to the other platforms. These findings underscore that platform-specific characteristics, such as optimisation techniques and library implementations, play a significant role in influencing Accuracy and F1-Score but have minimal impact on Recall and Precision.

Scikit-learn's superior recall and F1-Score, particularly when compared to KNIME, highlight the importance of selecting platforms that consistently demonstrate high sensitivity and specificity in critical applications like healthcare.

TABLE VII. TUKEY'S HSD TEST

	Accuracy			Recall			Precision			F1-Score		
Group 1 vs Group 2	Mean Difference	p-value	Significant	Mean Difference	p-value	Significant	Mean Difference	p-value	Significant	Mean Difference	p-value	Significant
KNIME vs MATLAB	0.0347	0.0987	No	0.023	0.6188	No	0.0523	0.1786	No	0.0383	0.1703	No
KNIME vs Scikit-learn	0.046	0.0302	Yes	0.0743	0.0311	Yes	0.0455	0.2557	No	0.06	0.0299	Yes
MATLAB vs Scikit-learn	0.0113	0.7345	No	0.0513	0.1368	No	-0.0068	0.9655	No	0.0218	0.52	No

It is important to highlight that this study does not aim to declare one platform superior to another in absolute terms. Instead, it provides critical insights into how platform architecture and design can influence ML model performance in different contexts. By investigating the inherent disparities in performance due to platform-specific characteristics, this work enables more informed decision-making when selecting platforms for predictive modelling. Ultimately, these findings contribute to a deeper understanding of how the interplay between ML algorithms and their implementation environments affects the reliability, accuracy, and effectiveness of models in real-world applications.

The scope of this study was intentionally focused on evaluating platform-dependent variations in ML classifier performance under default configurations. While this approach provides valuable insights, further research could extend these findings by incorporating additional analyses such as receiver operating characteristic (ROC) curves and precision-recall (PR) analyses. These techniques would enhance the interpretability of results, offering deeper insights into the trade-offs between sensitivity and specificity across platforms.

Another promising avenue for future exploration is classifier fusion, which could combine the strengths of individual classifiers to improve overall model performance. This technique holds potential for enhancing metrics such as accuracy, recall, and precision, especially in applications like cancer diagnosis, where both false positives and false negatives carry critical implications.

Expanding the study to include a broader range of machine learning algorithms, such as Support Vector Machines (SVM) and deep learning models, as well as additional datasets with varying characteristics, could further generalise the findings. Investigating the architectural differences of platforms, such as Scikit-learn, KNIME, and MATLAB, would also shed light on the underlying factors contributing to performance variations.

In conclusion, this study provides insights into how platform-specific characteristics influence ML model performance, offering practical guidance for platform selection in high-stakes applications like healthcare. While this research addresses a significant gap in the literature, it also lays the groundwork for further investigations into the interplay between ML algorithms and their implementation environments, enabling future advancements in predictive analytics and healthcare diagnostics.

REFERENCES

[1] A. S. Olowolayemo, A. Souag, and K. Sirlantzis,"Cancer: Investigating the impact of the implementation platform on machine learning models," The First International Conference on AI-Health (AIHealth 2024) IARIA, Mar. 2024, pp. 20-28, ISBN: 978-1-68558-136-7.

[2] B. S. Chhikara and K. Parang, "Global Cancer Statistics 2022: The trends projection analysis," Chem Biol Lett, vol. 10, pp. 451, Jan. 2023, Accessed: Dec. 01, 2024. [Online]. Available from: https://pubs.thesciencein.org/journal/index.php/cbl/article/view/451.

[3] "CANCER FACT SHEETS - Global Cancer Observatory." Accessed: Dec. 01, 2024. [Online]. Available from: https://gco.iarc.who.int/media/globocan/factsheets/cancers/39-all-cancers-fact-sheet.pdf.

[4] V. D. P. Jasti et al., "Computational technique based on machine learning and image processing for medical image analysis of Breast Cancer diagnosis," Security and Communication Networks, vol. 2022, pp.1-7, Mar. 2022, doi: 10.1155/2022/1918379.

[5] World Health Organization. "Cancer" Accessed: Dec. 01, 2024. [Online]. Available from: https://www.who.int/news-room/fact-sheets/detail/cancer.

[6] J. Boutry et al., "The evolution and ecology of benign tumors," Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, vol. 1877, pp. 188643, Jan. 2022, doi: 10.1016/j.bbcan.2021.188643.

[7] N. Behranvand et al., "Chemotherapy: a double-edged sword in cancer treatment," Cancer immunology, immunotherapy, vol. 71(3), pp. 507-526, Mar. 2022, doi: 10.1007/s00262-021-03013-3.

[8] J. Ko, M. M. Winslow, and J. Sage, "Mechanisms of small cell lung cancer metastasis," EMBO Mol Med, vol. 13, Jan. 2021, doi: 10.15252/emmm.202013122.

[9] S. U. Khan et al., "A machine learning-based approach for the segmentation and classification of malignant cells in breast cytology images using gray level co-occurrence matrix (GLCM) and support vector machine (SVM)," Neural Comp. and Applications, vol. 34, pp. 8365–8372, Jun. 2022, doi: 10.1007/s00521-021-05697-1.

[10] S. A. Alowais et al., "Revolutionizing healthcare: the role of artificial intelligence in clinical practice," BMC medical education, vol. 23, pp. 689, Dec. 2023, doi: 10.1186/s12909-023-04698-z.

[11] Y. Zhang et al., "Machine learning-based prognostic and metastasis models of kidney cancer," Cancer Innovation, vol. 1, pp. 124–134, Aug. 2022, doi: 10.1002/cai2.22.

[12] E. Y. Abbasi et al., "Optimising skin cancer survival prediction with ensemble techniques," Bioengineering, vol. 11, pp. 43, Dec. 2023, doi.org/10.3390/bioengineering11010043.

[13] R. Yang, I. F. Tsigelny, S. Kesari, and V. L. Kouznetsova, “Colorectal cancer detection via metabolites and machine learning,” Curr. Issues in Mol. Biology, vol. 46, pp. 4133–4146, May 2024, doi: 10.3390/cimb46050254.

[14] J.P. Villemin et al., "A cell-to-patient machine learning transfer approach uncovers novel basal-like breast cancer prognostic markers amongst alternative splice variants," BMC Biology, vol. 19, pp. 1-19, Apr. 2021, doi.org/10.1186/s12915-021-01002-7.

[15] K.A. Tran et al., "Deep learning in cancer diagnosis, prognosis and treatment selection,” Genome Medicine, vol. 13, pp. 1-17, Sept. 2021, doi.org/10.1186/s13073-021-00968-x.

[16] J. Kong et al., “Network-based machine learning approach to predict immunotherapy response in cancer patients,” Nature communications, vol. 13, pp. 1-15, Jun. 2022, doi: 10.1038/s41467-022-31535-6.

[17] W. Wolberg, O. Mangasarian, and W. Street, “Breast Cancer Wisconsin (Diagnostic),” UCI Machine Learning Repository, 1993, doi: 10.24432/C5DW2B.

[18] E. Michael, H. Ma, H. Li, and S. Qi, “An optimized framework for breast cancer classification using machine learning,” BioMed Research International, vol. 2022, pp. 8482022, Feb. 2022, doi: 10.1155/2022/8482022.

[19] S. Ara, A. Das, and A. Dey, "Malignant and benign breast cancer classification using machine learning algorithms," 2021 International Conference on Artificial Intelligence (ICAI), Islamabad, Pakistan, 2021, pp. 97-101, doi: 10.1109/ICAI52203.2021.9445249.

[20] M. Kumar, S. Singhal, S. Shekhar, B. Sharma, and G. Srivastava, “Optimized stacking ensemble learning model for breast cancer detection and classification using machine learning,” Sustainability, vol. 14, Jan. 2022, doi: 10.3390/su142113998.

[21] M. Ebrahim, A. A. H. Sedky, and S. Mesbah, “Accuracy assessment of machine learning algorithms used to predict breast cancer,” Data, vol. 8, Feb. 2023, doi: 10.3390/data8020035.

[22] C. Zhu, C. U. Idemudia, and W. Feng, “Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques,” Informatics in Medicine Unlocked, vol. 17, pp. 100179, Jan. 2019, doi: 10.1016/j.imu.2019.100179.

[23] A. K. M. Rahman, F. M. Shamrat, Z. Tasnim, J. Roy, and S. Hossain, “A comparative study on liver disease prediction using supervised machine learning algorithms,” International Journal of Scientific & Technology Research, vol. 8, Nov. 2019, pp. 419–422, ISSN 2277-8616.

[24] M. Minnoor and V. Baths, “Diagnosis of breast cancer using random forests,” Procedia Computer Science, vol. 218, pp. 429–437, Jan. 2023, doi: 10.1016/j.procs.2023.01.025.

[25] W. Li, Y. Yin, X. Quan, and H. Zhang, “Gene expression value prediction based on XGBoost algorithm,” Frontiers in Genetics, vol. 10, pp. 1077, Nov. 2019, doi: 10.3389/fgene.2019.01077.

[26] X. Wan, “Influence of feature scaling on convergence of gradient iterative algorithm,” Journal of physics: Conference series, vol. 1213, pp. 032021, Jun. 2019, doi: 10.1088/1742-6596/1213/3/032021.

[27] X. Yi et al., “Development and External Validation of Machine Learning-Based Models for Predicting Lung Metastasis in Kidney Cancer: A large population-based study,” International Journal of Clinical Practice, vol. 2023, pp. 1-13, Jun. 2023, doi: 10.1155/2023/8001899.

[28] D. A. Omondiagbe, S. Veeramani, and A. S. Sidhu, “Machine learning classification techniques for breast cancer diagnosis,” IOP Conference Series: Materials Science and Engineering, vol. 495, pp. 012033, Apr. 2019, doi: 10.1088/1757-899X/495/1/012033.

[29] R. Shafique et al., “Breast cancer prediction using fine needle aspiration features and upsampling with supervised machine learning,” Cancers, vol. 15, pp. 681, Jan. 2023, doi: 10.3390/cancers15030681.

[30] K. M. M. Uddin, N. Biswas, S. T. Rikta, and S. K. Dey, “Machine learning-based diagnosis of breast cancer utilizing feature optimisation technique,” Computer Methods and Programs in Biomedicine Update, vol. 3, pp. 100098, Jan. 2023, doi: 10.1016/j.cmpbup.2023.100098.

[31] T. Shamu et al., “Cancer incidence among people living with HIV in Zimbabwe: A record linkage study,” Cancer Reports, vol. 5, pp. e1597, 2022, doi: 10.1002/cnr2.1597.

[32] Q. T. N. Nguyen et al., “Machine learning approaches for predicting 5-year breast cancer survival: A multicenter study,” Cancer Science, vol. 114, pp. 4063–4072, Jul. 2023, doi: 10.1111/cas.15917.

[33] T. R. Mahesh et al., “Performance analysis of xGBoost ensemble methods for survivability with the classification of breast cancer,” Journal of Sensors, vol. 2022, pp. 4649510, Sep. 2022, doi: 10.1155/2022/4649510.

[34] Y. Zhang et al., “Machine learning-based prognostic and metastasis models of kidney cancer,” Cancer Innovation, vol. 1, pp. 124–134, Aug. 2022, doi: 10.1002/cai2.22.

[35] S. Aamir et al., “Predicting breast cancer leveraging supervised machine learning techniques,” Computational and Mathematical Methods in Medicine, vol. 2022, pp. 5869529, Aug. 2022, doi: 10.1155/2022/5869529.

[36] İ. Ateş and T. T. Bilgin, “The investigation of the success of different machine learning methods in breast cancer diagnosis,” Konuralp Medical Journal, vol. 13, pp. 347-356, Jun. 2021, doi: 10.18521/ktd.912462.

[37] L. Liu, “Research on logistic regression algorithm of breast cancer diagnose data by machine learning,” In 2018 International Conference on Robots & Intelligent System (ICRIS), May 2018, pp. 157–160, doi: 10.1109/ICRIS.2018.00049.

[38] X. Feng, Y. Cai, and R. Xin, “Optimising diabetes classification with a machine learning-based framework,” BMC Bioinformatics, vol. 24, pp. 428, Nov. 2023, doi: 10.1186/s12859-023-05467-x.

[39] S. Sumin, “The impact of Z-Score transformation scaling on the validity, reliability, and measurement error of instrument SATS-36,” JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia), vol. 11, pp. 166-180, Nov. 2022, dx.doi.org/10.15408/jp3i.v11i2.26591.

[40] M. Pagan, M. Zarlis, and A. Candra, “Investigating the impact of data scaling on the k-nearest neighbor algorithm,” Computer Science and Information Technologies, vol. 4, pp. 135-142, Jul. 2023, doi: 10.11591/csit. v4i2.p135-142.

[41] J. Cai, J. Luo, S. Wang, and S. Yang, “Feature selection in machine learning: A new perspective,” Neurocomputing, vol. 300, pp. 70–79, Jul. 2018, doi: 10.1016/j.neucom.2017.11.077.

[42] G. Alfian et al., “Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method,” Computers, vol. 11, pp. 136, Sep. 2022, doi: 10.3390/computers11090136.
[43] A.Olowolayemo (2023), Cancer3IPMLM GitHub. [Online]. Available from: https://github.com/ProfDee92/Cancer-3IPMLM/blob/main/README.md.
[44] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O’Sullivan, “A review of feature selection methods for machine learning-based disease risk prediction,” Frontiers in Bioinformatics, vol. 2, pp. 927312, Jun. 2022, doi.org/10.3389/fbinf.2022.927312.
[45] I. Kadhim Ajlan, H. Murad, A.A. Salim, and A. Fadhil Bin Yousif, “Extreme Learning machine algorithm for breast cancer diagnosis,” Multimedia Tools and Applications, pp. 1-20, Jun. 2024, doi.org/10.1007/s11042-024-19515-y.

Leveraging Large Language Models for the Identification of Human Emotional States

Clement Leung
School of Science and Engineering and
Guangdong Provincial Key Laboratory of
Future Networks of Intelligence
Chinese University of Hong Kong
Shenzhen, China
clementleung@cuhk.edu.cn

Zhifei Xu
School of Science and Engineering
Chinese University of Hong Kong
Shenzhen, China
zhifeixu1@link.cuhk.edu.cn

***Abstract*—This study explores emotion recognition, classification, and prediction, emphasizing its importance in safety-critical tasks where emotional competence can save lives. We classify emotions into positive (competent) and negative (incompetent) states and develop a stochastic model featuring an emotion stability factor, to measure how quickly individuals return to their baseline emotional state—lower indicates greater emotional stability. Our model provides a foundation for personalized emotional health strategies and tailored psychological treatments. Additionally, we evaluate ChatGPT-4's zero-shot capabilities in image-based sentiment reasoning compared to ResNet-50 and Vision Transformer models. Despite competitive performance, challenges such as unstable predictions underscore the complexity of mental health analysis in image conversations. We propose improvements through enhanced prompt engineering, model fine-tuning, and an ensemble approach combining each model's strengths to create a more accurate emotion classification system with significant implications for mental health applications.**

Keywords-image emotion recognition; large language model; zero-shot; emotion stability factor; ChatGPT4.

I. Introduction

The complex nature of human interactions makes it essential to accurately perceive and express emotions. Emotions shape individual experiences, influencing interactions, decision-making, and overall well-being. They help build connections and act as indicators of personal intentions and health.

For over a decade, research has focused on integrating emotional responsiveness into human-computer dialogue systems [1][2][3][4][5][6][7]. This field emphasizes the importance of understanding and predicting emotions for enhancing human-computer interactions and broader applications. As technology advances, there is a need to develop systems that can recognize and react to a wide range of emotions, ensuring natural interactions.

In modern society, individuals face stress from criticism, injustice, or relationship issues, often leading to emotional instability. Research shows that stress can result in harmful behaviors [8][9]. Such emotional distress can have severe consequences, from academic pressures leading to suicidal thoughts to road rage incidents. In high-stakes jobs like aviation or surgery, emotional regulation is critical for safety. Recent incidents, such as a pilot attempting to shut down an engine mid-flight due to depression, highlight the importance of emotional assessment. Key reasons why emotion recognition is important include:

- Understanding Emotions: Emotions guide how we perceive situations and decide what we want. Recognizing emotions helps individuals make choices that lead to positive interactions.
- Impact of Emotions: Emotions affect moods and behaviors, influencing relationships and well-being. Recognizing when we feel unhappy allows us to find solutions and regain control.
- Negative Emotions and Thoughts: Failing to recognize negative emotions can lead to harmful thoughts. By changing negative thinking patterns, individuals can reshape their perspectives and improve well-being.

This dissertation explores emotion recognition, classification, and prediction in high-stakes environments. We categorize emotional states into positive (competent) and negative (incompetent) groups. Our stochastic model analyzes how emotions evolve, using the emotional stability factor (λ) to quantify how quickly emotions return to a baseline. A lower λ indicates greater emotional stability, while a higher λ suggests frequent emotional fluctuations.

Understanding these dynamics can enhance therapeutic interventions, making them more effective and personalized. By anticipating emotional declines, the model enables the design of timely interventions. Knowing one's emotional stability factor is crucial for managing mental health, building resilience, and improving quality of life.

Section II reviews advancements in neural networks, including their use in online social networks and deep learning technologies. Models like ChatGPT [10] and Instruct-GPT [11] demonstrate the potential for improving emotion recognition.

This study investigates ChatGPT4's zero-shot capability to recognize emotions from images. While promising, ChatGPT4 has limitations, including prediction instability and reasoning inaccuracies. Enhancing its effectiveness through better prompt engineering and context selection is crucial for its application in mental health assessment. Our main contributions include:

- Applying mathematical models to predict emotional changes, offering a foundation for personalized mental health strategies.
- Using different emotional stability factors to understand how environments, individuals, or systems return to baseline states, aiding the design of targeted interventions.

- Exploring ChatGPT4's potential for emotion recognition in zero-shot scenarios.
- Developing conversational techniques in ChatGPT4 to enhance emotion recognition and discussing progress and limitations in its multimodal tasks [12][13][14].

The rest of this paper is organized as follows. Section II reviews the development of emotion recognition research and related studies. Section III explores the theoretical foundations of emotion recognition and prediction through rigorous mathematical logic and formulas. It lists the derivations of the algorithms and methods adopted in emotion recognition systems. By providing a solid mathematical framework, this section ensures that the subsequent empirical analysis is based on a robust theoretical model, which helps to gain a deeper understanding of the mechanisms behind emotion recognition technology. Section IV provides an experimental study focused on emotion recognition. It first details the selection process of appropriate datasets that are representative and comprehensive, thereby ensuring the validity and reliability of the experiments. This section further defines the specific tasks formulated as part of the research and outlines the goals and expected results of these experiments. The experimental results are analyzed, providing insights into the effectiveness of current emotion recognition models and highlighting potential areas for improvement. Section V and VI combine the experimental results with the capabilities of ChatGPT4 in the field of emotion recognition and prediction. It evaluates the progress brought by ChatGPT4 and discusses the performance of the model in the context of the experimental results. The paper ends with a reflection on the significance of these findings for future research and practical applications in this field, and provides suggestions for further research and enhancement of existing models.

II. Related work

What is emotion recognition? It refers to the technological process of identifying and interpreting human emotions. People naturally vary in their ability to accurately recognize others' feelings, which has led to the development of a specialized field that leverages technology to assist in this task. A key concept in this domain is affective forecasting, also known as hedonic forecasting. Affective forecasting involves predicting one's future emotional states, which plays a crucial role in shaping individual preferences, decisions, and behaviors. This interdisciplinary field, studied by both psychologists and economists, has a wide range of applications across various sectors.

The formal study of emotion theory began with Charles Darwin in 1872, establishing a foundation for future research into emotional expression. Building on Darwin's work, Paul Ekman introduced one of the most influential classification models in emotion recognition. He identified six basic emotions—joy, sadness, fear, anger, surprise, and disgust—which he proposed as universally recognizable across various cultures [15].

Robert Plutchik further expanded upon Ekman's model by developing the "wheel of emotions," which includes eight primary emotions: joy, trust, fear, surprise, sadness, disgust, anger, and anticipation [16]. Plutchik's wheel illustrates how these basic emotions can blend to form more complex emotional experiences, highlighting the dynamic nature of human emotions.

The theoretical frameworks for classifying emotions generally fall into two categories: categorical models and continuum models. Categorical models, or discrete emotion models, are based on the premise that there are a finite number of primary or basic emotions that are universally experienced, regardless of cultural or individual differences. In contrast, the continuum model views emotions as existing along a spectrum, capturing the nuances of emotional experience through dimensions such as valence (positive to negative), arousal (excited to calm), and dominance (sense of control or influence in a situation) [17] [18].

In recent years, emotion recognition and prediction technologies have developed into two primary types: single-modal and multimodal systems. Single-modal systems rely on a single data type, such as text, speech, or facial expressions, to detect and predict emotions [19] [20]. However, these systems often face limitations due to the constrained information provided by a single data source. For instance, relying solely on facial expressions may not fully capture an individual's emotional state, particularly in complex scenarios. The technology behind emotion recognition and prediction has gained significant attention due to its potential to enhance safety, support mental health, and improve user experiences. It has been identified as a vital factor in promoting human safety, making it a focus of extensive research [8][9]. The significance of emotion recognition and prediction is further highlighted by the growth of the Emotion Detection and Recognition (EDR) market. In 2024, the EDR market was valued at $57.25 billion and is projected to reach $139.44 billion by 2029, reflecting a remarkable compound annual growth rate (CAGR) of 19.49% from 2024 to 2029. This rapid expansion signifies the increasing demand and versatility of emotion recognition technology across various sectors. The growth is driven by the technology's ability to provide valuable insights into human emotions, proving invaluable in areas such as marketing, customer service, therapy, and security. By accurately detecting and responding to human emotions, this technology not only enhances interactions but also contributes to overall safety and well-being [21].

III. Methods

In real-world settings, emotions are dynamic and constantly shift from one state to another, which is crucial to consider in operational environments like workplace scheduling or hospital management. Accurately predicting individuals' emotional states is vital, especially when assigning tasks that demand high concentration and emotional stability. This need is even more pronounced in safety-critical roles. Workers experiencing emotional distress, whether due to unfair treatment or fatigue, may not only underperform but also pose safety risks to themselves and others. Thus, ensuring that employees, particularly in high-risk industries, maintain emotional stability is essential for workplace safety and efficiency.

Assessing emotional states is key to preventing potential accidents or errors that can arise from impaired judgment caused by negative emotions. For example, a worker struggling with unmanaged anger or severe sadness may lack the focus or decision-making capacity needed to safely operate machinery or make critical, split-second decisions. Therefore, understanding and managing employees' emotional well-being goes beyond enhancing individual performance; it is also about protecting the overall safety and smooth functioning of the workplace.

We apply Ekman's model, which identifies six primary emotions: happiness, sadness, fear, anger, surprise, and disgust. In this model, we categorize happiness and surprise as positive $+1$ emotions, suggesting a state of emotional well-being and openness. On the other hand, sadness, fear, anger, and disgust are categorized as negative -1 emotions, which might indicate potential emotional instability. Notably, in environments where safety is paramount, such as high-risk jobs, the categorization of surprise may shift. Typically considered a positive emotion due to its association with unexpected joy, surprise can be reclassified as negative in these critical contexts to ensure a conservative approach to emotional management, thereby reducing the risk of abrupt emotional reactions that could impair judgment or performance. Then, we propose that a crucial element in this model is the emotion stability factor, denoted by λ. Individuals with smaller emotion stability factor values exhibit more consistent emotional states. They are able to sustain either a positive or a neutral mood over extended periods, indicating resistance to sudden changes in emotions caused by external circumstances or internal reflections. Conversely, individuals with higher emotion stability factor values are prone to frequent emotional fluctuations. This heightened emotional reactivity can be attributed to various causes, including external influences like social interactions or internal factors such as hormonal variations. We will explore this concept in further detail in the sections that follow.

We represent the emotional state at time t by $S(t)$, where t denotes time and $S(t)$ describes the individual's emotional changes over time. The emotional state $S(t)$ can take on the following values:

$$S(t) = +1 \tag{1}$$

Equation (1) corresponds to the person's emotion being positive,

$$S(t) = -1 \tag{2}$$

Equation (2) means that the person's emotion is negative, corresponding to positive $+1$ and negative -1 emotions. Since various external happenings continually bombard humans, mood changes are often caused by events outside their control, possibly due to various factors. Such factors may be related to changing conditions of financial situation, relationships, health, work, stock market, and family, and the combination of these may cause a transition from a positive emotion state to a negative emotion state and vice versa.

First, let $S(0) = 1$. Then, we represent the transition time points by a Poisson Process. Now, $S(t) = 1$ if the number of transitions in the time interval $(0, t)$ is even, and $S(t) = -1$ if this number is odd. Therefore,

$$P[S(t) = 1|S(t) = 0] = p_0 + p_2 + p_4 + ... + ..., \tag{3}$$

where p_k is the number of Poisson points in $(0, t)$ with parameter λ:

$$p_k = e^{-\lambda t}\frac{(\lambda t)^k}{k!} \tag{4}$$

That is,

$$\begin{aligned} P[S(t) = 1|S(0) = 0] &= e^{-\lambda t}[1 + \frac{(\lambda t)^2}{2!} + \frac{(\lambda t)^4}{4!}... + ...] \\ &= e^{-\lambda t}\cosh \lambda t \end{aligned} \tag{5}$$

with

$$\cosh(t) = \frac{e^{-t} + e^{t}}{2} \tag{6}$$

$\cosh(t)$ can be expressed by its series expansion:

$$\cosh(t) = \sum_{k=0}^{\infty}\frac{t^{2k}}{(2k)!} \tag{7}$$

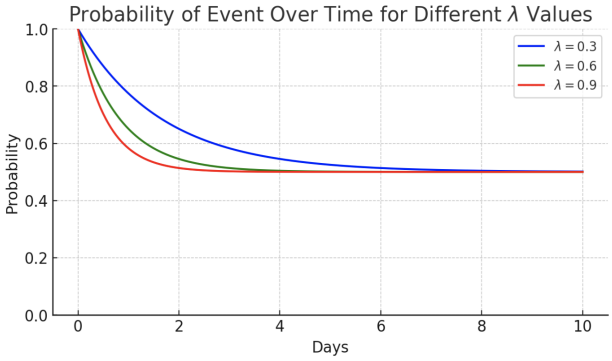


Figure 1. Different Emotion Stability Factors of the Equation (5)

On the other hand, Figure 1 provides a valuable perspective on how individuals might experience shifts in emotional states under different conditions. For example, individuals with a lower emotional stability factor, such as $\lambda = 0.3$, could be likened to those with a strong emotional foundation or support system, allowing them to maintain a higher probability of positive emotional states for longer periods. This may reflect traits like resilience, optimism, or effective coping mechanisms that slow the rate of emotional decay after positive experiences.

In contrast, higher emotional stability factor values, such as $\lambda = 0.9$, might represent individuals who are more susceptible to rapid emotional shifts due to external stimuli or internal factors, such as stress or a lack of coping resources. These individuals experience a quicker return to baseline emotional states, making them more vulnerable to negative emotional swings following positive experiences.

Examining the day-to-day probabilities reveals that achieving and maintaining a probability of over 50% for positive emotions

within the first day is feasible across all values of the emotional stability factor, with probabilities ranging from 60% to 80%. However, the duration over which this level is maintained varies significantly. While an emotional stability factor of $\lambda = 0.3$ can sustain this level for up to four days, an emotional stability factor of $\lambda = 0.9$ sees a rapid decrease, reaching 50% by the second day. This rapid decline may indicate the need for more frequent or stronger positive reinforcements or interventions to maintain positivity in such individuals.

Additionally, the stabilization of probabilities around the 50% mark from the second day for higher emotional stability factors and from the fourth day for lower factors suggests a natural equilibrium state in emotional dynamics. This equilibrium represents a critical point where the emotional state is equally likely to shift towards positive or negative, emphasizing the importance of timely psychological support or self-care practices during these periods to tilt the balance towards a more positive state.

In practical terms, this analysis can help tailor therapeutic approaches or wellness programs to fit individual emotional profiles. Understanding these dynamics could lead to more personalized and effective interventions, enhancing emotional well-being and stability by strategically addressing the decay rates of positive emotional states. Such insights are invaluable in clinical psychology, counseling, and personal development, where maintaining positive emotional states is essential for mental health and quality of life.

Now, if $S(t) = -1$ when the number of points in the time interval $(0, t)$ is odd, we have:

$$P[S(t) = -1|S(0) = 0]] = e^{-\lambda t}[1 + \frac{(\lambda t)^3}{3!} + \frac{(\lambda t)^5}{5!} ... + ...] = e^{-\lambda t} \sinh \lambda t \quad (8)$$

and

$$\sinh(t) = \frac{e^t - e^{-t}}{2} \quad (9)$$

This series is equivalent to the series expansion of $\sinh(t)$:

$$\sinh(t) = \sum_{k=0}^{\infty} \frac{t^{2k+1}}{(2k+1)!} \quad (10)$$

Equation (5) represents the probability that the emotion remains positive at time t, given that it was positive at time 0. Similarly, Equation (8) provides the probability that the emotion is positive at time t, given that it was negative at time 0. In both expressions, the emotional stability factor determines the rate of emotional change or decay over time. A larger value of this factor indicates more rapid emotional shifts, while a smaller value suggests slower changes.

Figure 2 illustrates the function $e^{-\lambda t} \sinh(\lambda t)$ with emotion stability factor values of 0.3, 0.6, and 0.9, offering insights into how probability changes over time, particularly in the context of emotional states or other processes that evolve or decay similarly. The graph shows that different values of the emotional stability factor (λ) can significantly affect the duration and intensity of these probability states over days.

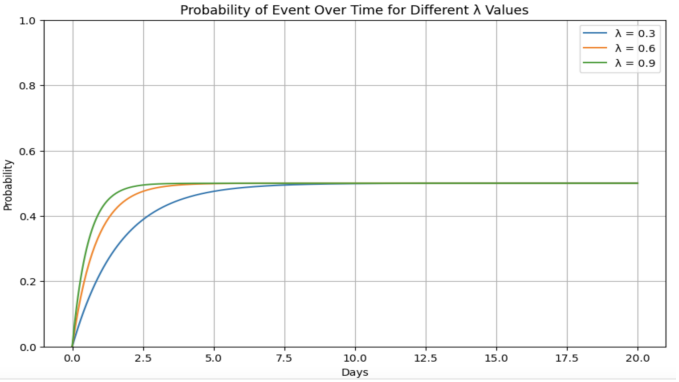


Figure 2. Different Emotion Stability Factors of the Equation (8)

For a lower decay constant, such as $\lambda = 0.3$, the probability starts strong and decays slowly, indicating a lingering effect. In the context of maintaining a negative emotional state, this suggests a higher likelihood of staying in that state for an extended period. Within the first day, the probability remains above 60%, indicating a persistent state that gradually stabilizes near 50% by the third day. This slower decay could reflect situations where the factors causing the emotional state are not quickly resolved.

When the decay constant increases to $\lambda = 0.6$, the probability declines more steeply, signifying a quicker dissipation of the state. While the probability is initially high, it drops below 60% by the end of the first day and approaches 50% by the second. This faster decay could represent scenarios where interventions are more effective or where individuals have better-coping mechanisms.

With a higher decay constant ($\lambda = 0.9$), the drop in probability is even more rapid. The steep descent shows that the state dissipates quickly, failing to sustain above 50% for more than two days and nearing this threshold by the end of the first day. This could represent situations where external support is highly effective, or the cause of the emotional state has a short-lived impact.

By analyzing these curves, we can infer the effectiveness of different strategies or inherent factors in managing specific emotional or physical states. The varying stability factor values represent different rates at which environments, individuals, or systems return to baseline or transition between states. Understanding these temporal dynamics is crucial in fields like psychology, where predicting the duration of a negative emotional state can inform tailored, time-sensitive interventions aligned with observed decay rates.

IV. Results

Emotion Recognition in Conversations (ERC) is extensively used in various contexts. This includes analyzing comments on social media platforms and monitoring personnel in high-stress environments. In addition, ERC technology is used in chatbots to accurately assess users' emotional states so they can tailor their responses accordingly. ChatGPT4 is one such conversational bot, as highlighted earlier. Within the context

of these interactions, we investigate and evaluate how well it recognizes and understands emotions and sentiments.

1) Dataset and Evaluation Graph: We using three different datasets from Kaggle, Facial Expressions Training Data [22], Emotion Detection [23], and Natural Human Face Images for Emotion Recognition [24].

Emotion Detection This dataset consists of 35,685 examples of 48x48 pixel grayscale images, which contain two folders, one is trained, and the other one is tested. The folders contain different categories of emotional images. In addition, the images have been labeled by the authors for different types of emotions, including anger, disgust, fear, happiness, neutral, sad, and surprise.

Facial Expressions Training Data AffectNet [25] is a large database of faces marked with "impact" (the psychological term for facial expressions). In order to accommodate common memory limitations in this dataset, the authors reduce the resolution to 96x96 for the neural network processing, which indicates that all images are 96x96 pixels. Meanwhile, using Singular Value Decomposition, each image's Principal Component Analysis is calculated. The threshold for the Percentage of the First Component (index 0) in the principal components (in short the PFC%) was set to lower than 90%. This means that most if not all of the monochromatic images were filtered out. Finally, the dataset is based on Affectnet-HQ, using a state-of-the-art Facial Expression Recognition (FER) model that refines the AffectNet original label to re-label its dataset, which contains eight emotional categories - anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise.

Natural Human Face Images for Emotion Recognition Since facial expression recognition is usually performed using standard datasets, such as the Facial Expression Recognition dataset (FER), Extended Cohn-Kanade dataset (CK+) and Karolinska Directed Emotional Faces dataset (KDEF) for machine learning, however, this dataset was collected from the internet and manually annotated to provide additional data on real faces, with over 5,500 + images with 8 emotions categories: anger, contempt, disgust, fear, happiness, neutrality, sadness and surprise. All images contain grayscale human faces (or sketches). Each image is 224 x 224 pixel grayscale in Portable Network Graphics (PNG) format. Images are sourced from the internet where they are freely available for download e.g., Google, Unsplash, Flickr, etc.

A. Task Definition

Based on the dataset description outlined in the previous section, we selected six emotions that are consistently annotated in each dataset for our experiments. These emotions include anger, disgust, happiness, neutrality, sadness, and surprise. Table I illustrates several examples of comparing annotations with ChatGPT4's predictions, with differences highlighted in red. For our experimental setup, we randomly selected 50 images representing six emotion types from each dataset and then submitted these images to ChatGPT4 for evaluation. Given that ChatGPT4 was launched in 2023, all of our experiments used this version of the model. At the same time, we chose to use the classic ResNet50 and ViT models as a comparison with ChatGPT.

ResNet50 is a classic deep convolutional neural network proposed by He Kaiming et al. [26], which effectively solves the gradient vanishing and degradation problems in deep networks by introducing residual connections. In the task of emotion recognition, subtle changes in facial expressions are crucial for accurate classification. With its powerful feature extraction capabilities, ResNet50 can effectively capture detailed features and complex patterns in face images. In our emotion recognition project, we used a pre-trained ResNet50 model as a feature extractor and then added a custom fully connected layer to adapt to the classification task of six emotion categories (anger, disgust, happiness, sadness, neutral, surprise). Through transfer learning, we made full use of the rich features learned by ResNet50 on large datasets such as ImageNet, which not only improved the accuracy of the model but also accelerated the training speed.

Vision Transformer (ViT) is a model proposed by the Google research team in 2020 that applies the Transformer architecture to image classification. Unlike traditional convolutional neural networks, ViT divides images into fixed-size image patches, which are then expanded into sequences and input into the Transformer network. ViT uses the self-attention mechanism to capture global dependencies in the image, which is very beneficial for understanding and classifying complex facial expressions. In our emotion recognition project, we took the pre-trained ViT model and fine-tuned it to adapt to the classification task of six emotion categories. The global feature capture capability of the ViT model gives it an advantage in identifying complex emotional expressions involving coordinated changes in multiple parts.

When evaluating the capabilities of ChatGPT4, we adopted a supervised learning approach and tested the performance of the model in a zero-shot prompting scenario designed specifically for this task. Each prediction of ChatGPT4 was carefully compared with our predefined cognitive assessment of the emotions depicted in the image. A correct prediction by ChatGPT4 that was consistent with our assessment was scored as 1, while a mismatch was scored as 0. In addition, each emotion was classified as positive, negative, or neutral according to ChatGPT4's description.

In addition, we generated a Receiver Operating Characteristic (ROC) curve based on the recorded results to quantify the accuracy of the model. When generating the ROC curve, emotions classified as positive (happy, neutral, or surprised) were marked with the factual result 1. In contrast, negative emotions (angry, disgusted, or sad) were marked with the factual result 0. The prediction accuracy of ChatGPT4 was then evaluated: if a positive emotion was correctly identified, it was recorded as 1; if not, it was recorded as 0. Similarly, for negative emotions, correct identification is recorded as 0, and incorrect identification is recorded as 1. In addition, the researchers also evaluated the confidence of each prediction using an evaluation

TABLE I
EXAMPLE OF CHATGPT4'S PREDICTION ON ERC TASK WITH IMAGES.

Image Content	Question	Annotation	Prediction
	What is the emotion of this person?	anger	surprise/shock/fear
	What is the emotion of this person?	happiness	happiness
	What is the emotion of this person?	happiness	happiness/joy
	What is the emotion of this person?	anger	frustration/concern/disapproval
	What is the emotion of this person?	sadness	sadness/crying
	What is the emotion of this person?	surprise	surprise

TABLE II
RESULT OF TRAINING RESNET50 MODEL PREDICTION ON SINGLE EMOTION RECOGNITION TASK WITH IMAGES.

loss	accuracy	val_loss	val_accuracy
2.2056	0.163	327522.7813	0.1597
1.9918	0.2430	2906.5945	0.1528
1.7945	0.2205	770.0659	0.1528
1.7411	0.2466	518.6124	0.1667
1.6758	0.2812	36.9275	0.1597
1.6480	0.2951	2.2501	0.2569
1.6283	0.2795	2.8927	0.1875
1.6717	0.2830	7.9509	0.1944
1.5971	0.3056	11.4519	0.1667
1.5564	0.3698	4.6194	0.1875
1.5738	0.3072	8.1097	0.2013

index of 1 to 3 points, where 1 indicates low confidence, 2 indicates medium confidence, and 3 indicates high confidence. This structured evaluation helps to quantitatively evaluate the effectiveness of ChatGPT4 in identifying and distinguishing various emotional states based on facial expressions.

B. Results

ResNet-50 is a well-known convolutional neural network (CNN) architecture designed to tackle image classification tasks efficiently.We can see the training result of ResNet-50 in the Table II. In this emotion recognition context, it achieves an accuracy of 30.72%, indicating a moderate ability to capture visual patterns relevant to different emotional expressions. This performance can be attributed to ResNet-50's strong feature extraction capability, which results from its deep convolutional layers and residual connections. These connections enable the network to learn complex features from images while mitigating the vanishing gradient problem, which commonly plagues deep networks during training. The training history of ResNet-50 shows a relatively faster convergence rate compared to ViT, suggesting it can achieve an acceptable level of accuracy within fewer epochs. This attribute makes it useful in situations where computational resources or training time is limited.

The Vision Transformer introduces a novel approach to image classification by adopting the self-attention mechanism, which has proven highly successful in natural language processing (NLP) tasks. Unlike CNNs, ViT processes images as a series of patches, learning relationships between different parts of the image. In this context, ViT achieves a lower accuracy of 19.96% which show that result of training ViT in the Table III, suggesting it encounters challenges in optimizing for emotion recognition. However, the training history reveals a more stable validation loss trajectory compared to ResNet-50, indicating consistent learning and potential for improvement.

Unlike ResNet-50, which excels in extracting local features, ViT captures global relationships across the entire image through its self-attention mechanism. This global context is

TABLE III
RESULT OF TRAINING VIT MODEL PREDICTION ON SINGLE EMOTION RECOGNITION TASK WITH IMAGES.

loss	accuracy	val_loss	val_accuracy
4.2872	0.1701	4.5549	0.1458
4.0688	0.1719	4.3315	0.1458
3.8694	0.1649	4.1312	0.1528
3.6925	0.1667	3.9536	0.1597
3.5350	0.1719	3.7932	0.1736
3.3960	0.1684	3.6490	0.1876
3.275	0.1667	3.5185	0.1667
3.1628	0.1649	3.3998	0.1806
3.0655	0.1667	3.2949	0.1806
2.9799	0.1719	3.1971	0.1806
2.9024	0.1701	3.1120	0.1806
2.8348	0.1788	3.0369	0.1806
2.7766	0.1788	2.9686	0.1806
2.7244	0.1927	2.9079	0.2014
2.6777	0.1910	2.8552	0.1945
2.6396	0.1910	2.8048	0.2084
2.6023	0.1910	2.763	0.2153
2.5710	0.1927	2.726	0.2153
2.5418	0.196	2.6929	0.2153
2.5176	0.1997	2.6595	0.2153

essential for emotion recognition, as emotions often depend on the overall configuration of facial features rather than isolated parts. The relatively stable validation loss in the training history of ViT suggests it learns in a more consistent manner. This stability could lead to better generalization with sufficient data and extended training. While the current accuracy is lower, its learning behavior indicates room for enhancement with more epochs, larger datasets, or additional fine-tuning. ViT can handle various input sizes and is not as constrained by fixed kernel sizes as CNNs, providing more flexibility when processing complex visual information such as varying facial expressions.

TABLE IV
RESULT OF CHATGPT4'S PREDICTION ON SINGLE EMOTION RECOGNITION TASK WITH IMAGES.

Emotion	**Accuracy**
Anger	30%
Disgust	19.30%
Happiness	78%
Neutral	69.34%
Sadness	44.30%
Surprise	70%

Table IV presents ChatGPT-4's performance in recognizing various single emotions. For the positive emotion of surprise, the model achieves an accuracy of approximately 70%. Meanwhile, the accuracy for identifying happiness reaches about 78%, highlighting ChatGPT-4's strong ability to detect positive emotions. When these two emotions are combined, the model maintains a commendable accuracy. As discussed in this section, both happiness and surprise correspond to a lower emotional stability factor, indicating that individuals can sustain these positive emotional states for extended periods.

In terms of negative emotions, ChatGPT-4's accuracy ranges from lowest to highest for disgust, anger, and sadness. During testing, we found that while the model can predict negative emotions in a zero-shot setting, it struggles to accurately differentiate between disgust and anger. The lowest accuracy is observed in identifying disgust, which may be attributed to the inconsistency in individual expressions of this emotion. Overall, ChatGPT-4's recognition accuracy from highest to lowest across the six emotions is as follows: happiness, surprise, neutral, fear, anger, and disgust. Notably, although ChatGPT-4 can correctly identify most images of surprise, it has difficulty determining whether the surprise is positive or negative, often categorizing it as a neutral emotion. This explains why the results for surprise closely resemble those for neutral.

As previously mentioned, mitigating the adverse effects of negative emotions is crucial, particularly for individuals in high-risk industries or groups. Therefore, our analysis focuses on three primary negative emotions: anger, disgust, and sadness, which are associated with a higher emotional stability factor and thus require greater attention. Our analysis reveals the following False Positive Rates (FPR): sadness at 0.3267, anger at 0.4800, and disgust at 0.6467. These FPR values indicate that disgust is the most challenging emotion to identify accurately, making it the most difficult category among the six evaluated emotions. The overall accuracy in detecting negative emotions remains insufficient. To address this, enhancing the prompts provided to ChatGPT-4 is essential. Although the model can recognize the presence of negative emotions in a zero-shot scenario, it struggles to accurately distinguish between specific states such as disgust, contempt, or anger. Therefore, more refined prompt engineering could improve the model's ability to discern these nuanced emotional states.

V. Discussion and Evaluation

The training history indicates that ResNet-50 starts with high and unstable validation loss, hinting at a potential overfitting risk. This suggests sensitivity to noise in the data, which may require careful hyperparameter tuning, regularization, and augmentation techniques to improve generalization. While its 30.72% accuracy shows some effectiveness, it remains significantly lower than GPT-4's 50.99%. This lower performance indicates that ResNet-50 might struggle to capture the holistic context of emotions, focusing instead on local features without considering global facial patterns.

The initial high training loss and the current low accuracy of 19.96% suggest that ViT requires more epochs and possibly larger datasets to reach optimal performance. This slower convergence is often a trade-off for its global feature extraction capability. ViT typically benefits from vast amounts of training data to learn complex patterns effectively. When applied to smaller datasets, such as those typically available for emotion recognition, it may struggle to outperform more established CNNs like ResNet-50 unless supplemented with data augmentation or transfer learning strategies.

During training, we often encounter inconsistencies between the images in specific datasets and our real-life perceptions. Since individuals react differently to images, biases can inadvertently be introduced into the emotional recognition of specific photos. For these particular images, we rely on human judgment as the ultimate arbiter, comparing our assessments with the outputs from ChatGPT4 to identify discrepancies and possible biases.

Furthermore, a significant challenge arises from the misalignment between the predictions of ChatGPT4 and the guidelines provided in the dataset. This divergence highlights a fundamental issue where ChatGPT4 often deviates from the dataset norms. For instance, if the dataset annotates an image depicting anger based on the subject's facial expression, ChatGPT4 might interpret the same expression as sadness or confusion. This difference does not necessarily imply that one interpretation is correct and the other erroneous; instead, it underscores the variability in applying different emotional criteria, both falling under the umbrella of negative emotions.

Upon deeper analysis, the discrepancy in interpretation may not solely be a flaw in ChatGPT4's functionality but could also stem from inadequate prompt design. As we incorporate more complex prompt word guides, it becomes increasingly challenging to encompass the nuanced emotional contexts with a limited set of instructions. This situation opens up avenues for future improvements: if strict adherence to the dataset's guidelines is not mandatory, enhancing the model based on broader prompt settings (such as more descriptive cues about people in images) could be a viable strategy. However, reliance on dataset labels for evaluation may be less effective, potentially necessitating a more comprehensive manual review process. Conversely, if absolute fidelity to the dataset's guidelines is essential, more than employing a few general prompt settings may be required. Instead, a more structured and supervised fine-tuning of the model may be necessary to ensure accuracy and adherence to specific emotional classifications.

VI. Conclusion and Future Work

In this paper, we studied the problem of emotion recognition, classification, and prediction. Emotion recognition situations have become increasingly important due to the role played by emotions in key personnel in how they perform their duties, which, particularly for safety-critical tasks, can mean jeopardizing or saving lives. To determine emotional competence in discharging duties, we group different emotion states into positive (i.e., competent) or negative (i.e., incompetent). We have developed a stochastic model to understand emotional dynamics and the evolution of emotional states. Our model is

able to represent the influences on how individuals experience and sustain their emotions over time and offers a structured way to investigate the intricacies of human emotions. By assigning numerical values to emotional decay rates, it is possible to quantify how quickly individuals return to a baseline emotional state after experiencing fluctuations. This approach allows for the observation and comparison of emotional stability across different individuals over periods.

A key parameter in this model is the emotion stability factor λ. Individuals with lower emotion stability factor values tend to have more stable emotional states. They can maintain a positive or neutral emotional condition for a longer duration, which suggests resilience to rapid shifts in mood due to external events or internal thoughts. Conversely, those with higher emotion stability factor values are more susceptible to emotional swings. Such sensitivity might be due to various factors, including external stimuli like social interactions or internal changes such as hormonal shifts.

In summary, using mathematical models to analyze and interpret the evolution of emotions based on emotion stability factor values provides a scientific basis for customized emotional health strategies. This approach enhances our understanding of emotional dynamics and supports the development of more effective psychological treatments and wellness programs tailored to individual needs. Such methodologies could lead to significant advancements in mental health practices, ultimately improving the quality of life for various populations.

Additionally, through experiments, we delve into the zero-shot capabilities of ChatGPT4 for image-based sentiment reasoning and judgment, comparing results with ResNet-50 and ViT models. Results show that while ChatGPT4's predictive power holds up well against the other two models, there is still much room for improvement, primarily by integrating mental health analysis and humanistic inputs. The main challenges identified include unstable predictions and inaccurate reasoning. Our results highlight the inherent difficulty of tasks such as mental health analysis and sentiment reasoning for image conversations, which remain daunting tasks for ChatGPT. However, we believe that further progress can be made to enhance the performance of ChatGPT4 through improved prompt engineering and more selective integration of contextual examples. Such enhancements are critical for their potential applications in real-world mental health settings and other related fields, where nuanced sentiment understanding is crucial.

In future work, we may fine-tune the ResNet-50 and ViT models to improve their generalization capabilities. For ResNet-50, this may involve using advanced regularization techniques, data augmentation, and carefully tuning hyperparameters to stabilize the training process. For ViT, increasing the number of training epochs, using transfer learning from larger datasets, and employing data augmentation may improve its performance. In addition, an ensemble approach that combines the predictions of ResNet-50, ViT, and GPT-4 may help. The ensemble can use ResNet-50 for detailed local feature extraction, ViT to capture global context, and GPT-4 to integrate external context and provide a multimodal perspective. By weighting the predictions according to each model's respective strengths, the ensemble can produce a more comprehensive and accurate emotion classification system.

References

[1] C. H. C. Leung and Z. Xu, "Emotional recognition and classification using large language models", in *The Ninth International Conference on Neuroscience and Cognitive Brain Information, BRAININFO*, ThinkMind, 2024, pp. 4–10.

[2] C. H. C. Leung, J. J. Deng, and Y. Li, "Enhanced human-machine interactive learning for multimodal emotion recognition in dialogue system", in *Proceedings of the 2022 5th International Conference on Algorithms, Computing and Artificial Intelligence*, 2022, pp. 1–7.

[3] J. J. Deng and C. H. C. Leung, "Towards learning a joint representation from transformer in multimodal emotion recognition", in *Brain Informatics: 14th International Conference, BI 2021, Virtual Event, September 17–19, 2021, Proceedings 14*, Springer, 2021, pp. 179–188.

[4] J. J. Deng, C. H. C. Leung, and Y. Li, "Multimodal emotion recognition using transfer learning on audio and text data", in *Computational Science and Its Applications–ICCSA 2021: 21st International Conference, Cagliari, Italy, September 13–16, 2021, Proceedings, Part III 21*, Springer, 2021, pp. 552–563.

[5] J. J. Deng and C. H. C. Leung, "Deep convolutional and recurrent neural networks for emotion recognition from human behaviors", in *Computational Science and Its Applications–ICCSA 2020: 20th International Conference, Cagliari, Italy, July 1–4, 2020, Proceedings, Part II 20*, Springer, 2020, pp. 550–561.

[6] J. J. Deng and C. H. C. Leung, "Dynamic time warping for music retrieval using time series modeling of musical emotions", *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 137–151, 2015.

[7] J. J. Deng, C. H. C. Leung, A. Milani, and L. Chen, "Emotional states associated with music: Classification, prediction of changes, and consideration in recommendation", *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 5, no. 1, pp. 1–36, 2015.

[8] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: A narrative review", *NPJ Digital Medicine*, vol. 5, no. 1, p. 46, 2022.

[9] D. Ciraolo *et al.*, "Emotional artificial intelligence enabled facial expression recognition for tele-rehabilitation: A preliminary study", in *2023 IEEE Symposium on Computers and Communications (ISCC)*, IEEE, 2023, pp. 1–6.

[10] B. Mann *et al.*, "Language models are few-shot learners", *arXiv preprint arXiv:2005.14165*, 2020.

[11] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback", *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

[12] K. Yang, S. Ji, T. Zhang, Q. Xie, and S. Ananiadou, "On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis", *arXiv preprint arXiv:2304.03347*, 2023.

[13] W. Zhao *et al.*, "Is chatgpt equipped with emotional dialogue capabilities?", *arXiv preprint arXiv:2304.09582*, 2023.

[14] H.-D. Le, G.-S. Lee, S.-H. Kim, S. Kim, and H.-J. Yang, "Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning", *IEEE Access*, vol. 11, pp. 14 742–14 751, 2023.

[15] P. Ekman, *Facial expressions of emotion: New findings, new questions*, 1992.

[16] P. Robert, *Emotion: Theory, research, and experience. vol. 1: Theories of emotion*, 1980.

[17] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion recognition in context", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1667–1675.

[18] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Context based emotion recognition using emotic dataset", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 11, pp. 2755–2766, 2019.

[19] Z. Lian, Y. Li, J.-H. Tao, J. Huang, and M.-Y. Niu, "Expression analysis based on face regions in real-world conditions", *International Journal of Automation and Computing*, vol. 17, pp. 96–107, 2020.

[20] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks", *Biomedical Signal Processing and Control*, vol. 59, p. 101 894, 2020.

[21] G. Trends and Forecasts, "Emotion detection and recognition market size and share analysis", 2024, [Online]. Available: https://www.mordorintelligence.com/industry-reports/emotion-detection-and-recognition-edr-market.

[22] N. Segal, *Facial Expressions Training Data*, https://www.kaggle.com/datasets/noamsegal/affectnet-training-data, 2022.

[23] ananthu017, *Emotion Dection*, https://www.kaggle.com/datasets/ananthu017/emotion-detection-fer, 2020.

[24] S. Vaidya, *Natural Human Face Images for Emotion Recognition*, https://www.kaggle.com/datasets/sudarshanvaidya/random-images-for-face-emotion-recognition, 2020.

[25] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild", *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[26] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2015. arXiv: 1512.03385 [cs.CV].

Connotation and 3D Modeling from Limited, Raw Textual Descriptions

Ella Berman
Computer Science
Grinnell College
Grinnell, USA
email: bermanel@grinnell.edu

Mahiro Noda
Biological Chemistry
Grinnell College
Grinnell, USA
email: nodamahi@grinnell.edu

Kailee Shermak
Sociology
Grinnell College
Grinnell, USA
email: shermakk@grinnell.edu

Zi Ye
Computer Science
Grinnell College
Grinnell, USA
email: yezi@grinnell.edu

David Rothfusz
Computer Science
Grinnell College
Grinnell, USA
email: rothfusz2@grinnell.edu

Jiayi Chen
Risk Management
Pennsylvania State University
State College, USA
email: jjc7655@psu.edu

Thammik Leungpathomaram
Computer Science
Grinnell College
Grinnell, USA
email: leungpat@grinnell.edu

Shuta Shibue
Computer Science
Grinnell College
Grinnell, USA
email: shibuesh@grinnell.edu

Chenxing Liu
Computer Science
Grinnell College
Grinnell, USA
email: liutommy@grinnell.edu

Fernanda Eliott
Computer Science
Grinnell College
Grinnell, USA
email: eliottfe@grinnell.edu

***Abstract*—In emotion-rich contexts, how do you comprehend the meaning behind your perception? This exploratory multi-phase project seeks to gather insights into how abstraction and emotions travel different spaces. The investigated spaces are: images, human-made textual descriptions, mental models, and 3D (three-dimensional) scenes. In previous work, we described our project idea; here, we detail our pilot for Project Phases 1 and 2, in which a team first creates *raw descriptions* of memes (in addition to creating detailed descriptions and the Observer-Centered Dataset Attributes) so that the Phase 2 team, so-called modelers, read the raw descriptions and build a 3D scene as accurately and faithfully as possible to the meaning behind their perception of the description. Raw descriptions are created by "unsaying" (*i.e.*, by identifying and removing the *unsaid elements* from a detailed description); and therefore, are more vague than alt-text since raw description purposefully leave details out (to see if the modelers "got" the message in spite of gaps). We designed a diagram to illustrate how modelers decided a 3D scene was complete, called "*Accuracy* and *Faithfulness* Gateways Diagram", detailed here. We launched this project as a pilot to inform our methods to ensure objectivity and replicability. A key challenge in identifying the *unsaid elements* comes from making the implicit explicit, and our approach to accomplishing that can inspire frameworks for detecting biases and microaggressions in visual content and help to create cultural sensitivity awareness. We pinpoint our work's social impact applications, which will be detailed in future work. Finally, investigating abstraction within and across spaces is notably relevant right now. In fact, as more people interact with generative AI platforms (such as AutoGen or Vertex AI), prompt designers deal with and add abstraction into a prompt as they instruct an AI-powered model to behave in certain ways.**

***Keywords-abstraction; connotation; memes; 3D-modeling; textual descriptions*.**

I. Introduction

Connotation offers versatile approaches to communication. It can support argot languages or even hidden messages and expression against oppression, such as in Brazilian songs known for their response to dictatorship, e.g., "Sinal Fechado" (Paulinho da Viola, 1969), "Comportamento Geral" (Gonzaguinha, 1973), "Mosca na Sopa" (Raul Seixas, 1973), and "Cálice" (Chico Buarque and Gilberto Gil, 1978). These songs illustrate how abstraction, emotions, and connotation blended together can help to create complex messages.

Besides songs, poems, and others, memes widely shared online rely on connotation to deliver a message. Here, we build on our previous work [1] on memes [2], which are a "form of media communicating a thought or idea through some shared understanding" [3]. Memes "often hide complex, abstract reasoning mechanisms behind their humorous front" [3]. Connotative meanings refer to the "associations, overtones, and feel that a concept has, rather than what it refers to explicitly (or denotes, hence denotative meaning). Two words with the same reference or definition may have different connotations" [4]. "In writing, you can choose a word that has a clear denotation and few connotations—a word like tall or quiet—or you can choose a word that connotes something more—like statuesque or tranquil" [5].

But how do we *comprehend* the overtones and overall meaning behind our perception? More importantly, how do we play with connotation to create hidden messages that others can understand? Emotion knowledge enables children

to identify emotions in themselves and others and facilitates emotion recognition in complex social situations. Thus, social-cognitive processes, such as theory of mind (ToM), may contribute to developing emotion knowledge by helping children comprehend the emotion expression's variability across individuals and situations [6]. Theory of mind can be defined as the "human ability to ascribe mental states, intentions, and feelings to other human agents and to oneself" [7].

When telling a story, we do not provide every single detail; we expect others to fill in the gaps and evoke mental models consistent with the story. E.g., if the story involves a library, it may be associated with a quiet place filled with books and other associated behaviors/rules/objects (if those clues correspond to one's cultural experiences). Mental models are "internal representations of the external world consisting of causal beliefs that help individuals deduce what will happen in a particular situation" [8]. Meanwhile, emotional mental models cover emotions and feelings connected to mental models: "Mental models cause certain expectations/thoughts of how things should look like/work and connect certain emotions with this. Consequently, a mental model is a cognitive and an emotional framework in the brain, influenced by person's personality (genes) and the environment including social variables" [9] (see [10] for a theoretical review on the role of shared mental models in human-AI teams).

Yet, what if you wanted to somehow architect similar skills into a machine? Suppose you wanted to model an *emotion-driven Artificial Intelligence* (AI) system able to cooperate purposefully with humans and other biological creatures. You may ask: "How to encode abstract and emotion-rich contexts into an AI system's mental models and assist its decision-making process?" That is one of our main research goals, although we wonder: 1. Would that enable a more holistic contextual evaluation and better-informed AI's decision-making process? and 2. If one is to architect an AI system modeled after emotions and feelings, should it be influenced in any way by task-irrelevant emotion stimuli [11]? If so, what does that look like?

We use the term *emotion-rich* to convey emotional messages that the human senses can perceive (which could be translated, in robotics, for example, *via* the robot's sensors [12]). Our lab builds cognitively inspired computational models, and we are designing a computational architecture that uses traditional Reinforcement Learning techniques (RL) [13] and models emotions and moral processes [14], [15], [16]. RL is "learning what to do—how to map situations to actions – so as to maximize a numerical reward signal. The learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them. In the most interesting and challenging cases, actions may affect not only the immediate reward but also the next situation and, through that, all subsequent rewards. These two characteristics—trial-and-error search and delayed reward—are the two most important distinguishing features of reinforcement learning" [13].

To tackle that goal, we decided to examine humans first and narrowed our questions to "How do abstraction and emotions travel different spaces?" In [1], we present our multi-phase multi-team project idea to investigate that question, which explores distinct spaces: images of memes, human-made textual descriptions, mental models, and 3D scenes (see figs. 1 and 3).

A short overview of the project's Phases 0-2 is as follows: **Phase 0.** Manually collect images of memes. **Phase 1.** Analyze the images and create a database with 1. raw textual image descriptions, 2. detailed textual image descriptions, and 3. a set of attributes to analyze the memes, resulting in 4. the Observer-Centered Dataset. **Phase 2.** Create 3D scenes from the raw descriptions. We launched this project as a pilot, enabling us to create methods and gateways across and within phases using an *ad-hoc* and data-driven approach. Therefore, we will rerun the project once all phases are consolidated.

We hypothesize that by investigating how abstraction and emotions travel different spaces, we will gather insights into key elements for producing a consistent and holistic understanding of complex, abstract messages and connotations – finally, getting a better picture of how and what to model in an *emotion-driven* AI system that uses traditional RL techniques.

"Abstraction enables humans to distill a cascade of sensory experiences into a useful format for making sense of the world and generalizing to new contexts" [17]. Highlighting that knowledge exists at multiple levels of abstraction, Reed [18] provides a taxonomic analysis of abstraction that examines three senses of abstraction: "(a) an abstract entity is a concept that has no material referent, (b) abstraction focuses on only some attributes of multicomponent stimuli, and (c) an abstract idea applies to many particular instances of a category." Forward, Ho et al. [19] illustrate the importance of abstraction for AI and RL frameworks: abstractions are important for adaptive decision-making, e.g., abstractions guide exploration and generalization, facilitate efficient trade-offs, and simplify computation. Note that providing AI modeling details and identifying different uses of abstraction (e.g., visual abstraction, relational abstraction, temporal abstraction) falls out of our scope; the same goes for disassociating abstraction from emotions and connotation, but we provide definitions in the Glossary, see Appendix A.

We use "abstraction" as a blanket term for *something untied from concrete elements*, which covers both abstract words (e.g., *honor* and *freedom*) and/or dealing with abstraction (abstract problem-solving)."There is reason to assume that abstract concepts are more sensitive to contextual constraints than concrete concepts" [20] and "Statistically, abstract words are more emotionally valenced than are concrete words" [21]. Finally, challenges from interpreting an image that poses multiple emotional mental models drove us to *networked emotions* [22], helping us deal with the messy layers of emotions in meme comprehension (see Section VI). Still, we understand that humor "is a universal phenomenon but is also culturally tinted", and "some humor coping strategies may have different connotations under different cultural backgrounds, which would directly impact how humor is used in different cultural backgrounds" [23].

Hence, in our research project, abstraction intercepts emotions to the extent that, similar to the *Telephone Game*, people form different mental/emotional models as a message travels through them – and aspects of comprehending a message from a raw description are abstract and open-ended. (The Telephone Game starts with a line, or circle, of people; the first person in line privately receives or creates a message and whispers it to the next person in line. The process repeats until the end of the line; finally, the last person shares the message out loud to check if the original message accurately made its way through the whispers.) We acknowledge this research's challenges and limitations: abstraction and emotions are multifaceted topics, whose combination with technologies brings even more layers. Still, in spite of the challenges, our research outcomes motivate directions for social impact tools (see Section V), in addition to insights for cognitively-inspired AI modeling and dealing with networked emotions.

A note on Phase 0 image collection. We identified the digital space as a good fit for our purposes given the emotions' social nature and their central place in digital cultures: "The socially mediated communication of emotion is intricately linked to the social textures of networking technologies" [24]. This led us to images that convey jokes or metaphors characteristic of memes since they often hide abstract reasoning mechanisms and given their ability to be either easily understood or learned through examples, making them a viable format for idea transfer [3]. Memes can be used for various purposes, e.g., to entertain, instruct, or express political views and expose others to political content [25]; for simplicity, we target their use within humor or entertainment.

Our contributions are to:

1) Detail a methodological breakdown to textually describe memes (or similar images). Albeit the raw descriptions' purpose is to check what message will be encoded by the 3D modelers, our methods are still relevant to others working with textual description tools. We provide two kinds of image descriptions: detailed and raw (the latter is created *via* the identification and removal of *unsaid elements* from detailed descriptions).
2) Visually organize and contextualize a set of attributes to inform the analysis of memes. The attributes take into account the observer's perspective and networked emotions; we call those the Observer-Centered Dataset attributes.
3) Illustrate how a 3D modeler deals with limited, raw textual descriptions and decides whether a 3D scene is complete. Two gateways (*accuracy* and *faithfulness*) are identified, and they serve as a checkpoint for evaluating and inspecting the 3D model before it is complete, becoming a so-called 3D scene – modelers decide whether the scene sufficiently reflects the observable visual features of the mental models they created based on the raw description.
4) Provide a glossary and a multidisciplinary literature review as we situate our research.

Although it is not our claim that our Gateways Diagram covers the 3D modeling process in general, we do hope this work can a) benefit the decision-making process of similar 3D modeling initiatives and b) inform the creation of richer alt-text tools or even other assistive technologies, such as 3D modeling tools for the visually impaired – ultimately assisting in creating 3D printing blueprints; see resources such as Round Table on Information Access for People with Print Disabilities [26], See3D [27], and the Accessible Graphics hub [28]. More specifically, we hope to inform richer assistive technology tools' creation as we call attention to a tension between the *unsaid elements* (in a description or explanation, for example) and the audience's *assumed elements*. Therefore, in how abstraction and emotions make their way through different spaces, especially when a message is heavy on connotation.

That tension, worked through our proposed descriptions' breakdown, dataset, and diagram, can **inform the development of AI tools better equipped to deal with abstraction** (e.g., using Generative AI tools for creating a 3D scene from vague prompts), connotations, and cultural elements, aiming for culturally sensitive human-machine interaction and output/outcomes. Deviating from memes, one may ponder upon the everyday news on AI achievements, which play with connotations and anthropomorphism [29].

This work is organized as follows: introduction in Section I, followed by our methods, which split into two: Project's Phase 1 details in Section II, followed by Phase 2 in Section III. We show our research outcomes, such as the Observer-Centered Dataset attributes and a sample of 3D scenes' static images in Section IV, discussion in Section V, followed by related literature in Section VI, and conclusion in Section VII.

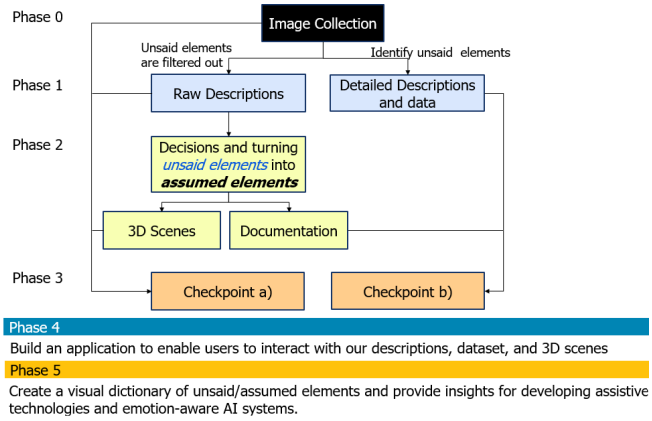


Fig. 1. Project Overview. Project phases 1 and 2 are this manuscript's focus: producing and encoding raw textual descriptions into a 3D scene (and documenting the decision-making process).

II. Background and Methods

A summary of all project phases is given next, and an overview in Figure 1. As Figure 2 shows, Phase 1 covers human-made textual descriptions followed by the memes' attributes identification; then, Figure 3 illustrates, for Phases

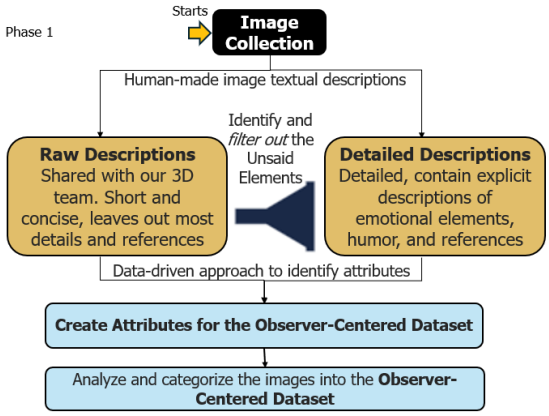


Fig. 2. Phase 1 starts by accessing Phase 0 meme collection to create detailed and raw descriptions and the Observer-Centered Dataset attributes. Finally, the memes are categorized into the dataset.

$0-3$, in what ways we envision abstraction and emotions traveling spaces.

Note that **each phase has a unique team with a strict non-sharing policy**: everything a team produces is kept within the team only, unless at the Phase's cycle conclusion when we share our results with the community – for consistency, we use "we" across this manuscript as we integrate the teams' results. We launched this project as a *pilot* to establish procedures and methods to ensure objectivity and replicability. To that end, we investigated related literature combined with a data-driven approach (see Section IV-C).

Project Phases in a Nutshell:

- Phase 0, **Image (Meme) Collection**. Manually collect images that hide complex, abstract reasoning mechanisms characteristic in memes – any political or hateful content is forbidden. The ≈ 400 memes were collected from social platforms such as Instagram and WhatsApp, and they cover two countries (the USA and Brazil), as we sought to investigate more than one country. As expected, humorous memes are abundant online; therefore, our collection sits within humor, with just a few exceptions (see Figure 7). Still, some memes seek to evoke humor from negative tones, given the use of self-deprecating humor.
- Phase 1, **Database**. Write raw and detailed image descriptions in English and categorize the memes in a dataset called the Observer-Centered Dataset. Feed the *Phase 2* team with raw descriptions – i.e., leaving details out, by which we named *unsaid elements*. Example of a raw description: *A* soaking wet cat sits inside a sink with open eyes that pop out. There is a leading text: "I leave the bathroom shaking cold, and the person asks:" follow-up text: "Are you cold?' Nope, a ghost is entering me."
- Phase 2, **3D Scenes and Decisions**. Without access to the memes, interpret and encode the raw image descriptions into a 3D scene using a tool such as Blender and document the decision-making process. *Unsaid elements* can either be on the a) concrete side, e.g., it mentions a cat on a sink but no details about the fur's color or the sink's shape, size, and material/color, or the environment; or b) more abstract and emotionally-tinted, e.g., 3D modelers may reflect: "this seems to imply discomfort; is it supposed to be humorous?" Hence, 3D modelers have to fill in the gaps and make decisions to build a 3D scene, by which we call *assumed elements*. Therefore, *unsaid elements* from Phase 1 become *missing elements* in Phase 2, as modelers identify that something is missing in the description and subsequently make assumptions of how to fill in the gaps, resulting in the *assumed elements* – see Figure 3.
- Phase 3, **Checkpoint**. Compare: a) raw descriptions, memes, and 3D scenes (focus on the 3D scenes' canonical view, which should coincide with the front view), and b) unsaid with *assumed elements* and documentation. Examine how/if those differ, analyze our dataset, and document what we learned about abstraction/emotions across spaces.
- Phase 4, **Software application**. We will apply human-centered design (HCD) practices to develop a software application, e.g., Shiny app [30], to enable people to interact with our project's data and outcomes.
- In Phase 5, we will investigate in what ways our findings can inform the development of a **Visual Dictionary** that refers back to emotions, abstraction, and connotative meanings – we hypothesize this work will provide valuable insights for fostering assistive technologies and modeling *emotion-driven* AI systems.

To recap, our overall goal is to architect an *emotion-driven* AI system; to inform our processes, we are investigating how abstraction and emotions travel through spaces. The comparison *Unsaid Elements* (from Phase 1) with the *Missing* and *Assumed Elements* (from Phase 2) will be key in investigating/mapping the different elements people may combine to interpret abstract messages. Likewise, our dataset will enable us to filter and group meme details in various ways (such as comparing memes from Brazil and the US), contributing insights for AI modeling, such as dealing with humor and connotation in different cultures. We detail Phase 1 next.

A. Textual Descriptions and a Narrative Approach

Phase 1's overall goal is to devise a method for creating descriptions that are *as raw as possible* but still retain the meme's overall meaning. Abstract and emotionally-tinted elements are a) included in a detailed description but b) filtered out from a raw description. Therefore, it involves dealing with both connotative and denotative meanings. According to Schnotz [31], "text comprehension includes the formation of at least three kinds of mental representations: a text surface representation, a propositional representation, and a mental model" and "Inferences are an integral component of text comprehension, because the author of a text omits information which can be easily completed by the reader." More than that, through the *unsaid elements*, we seek to identify and omit more information than one typically would to convey a message.

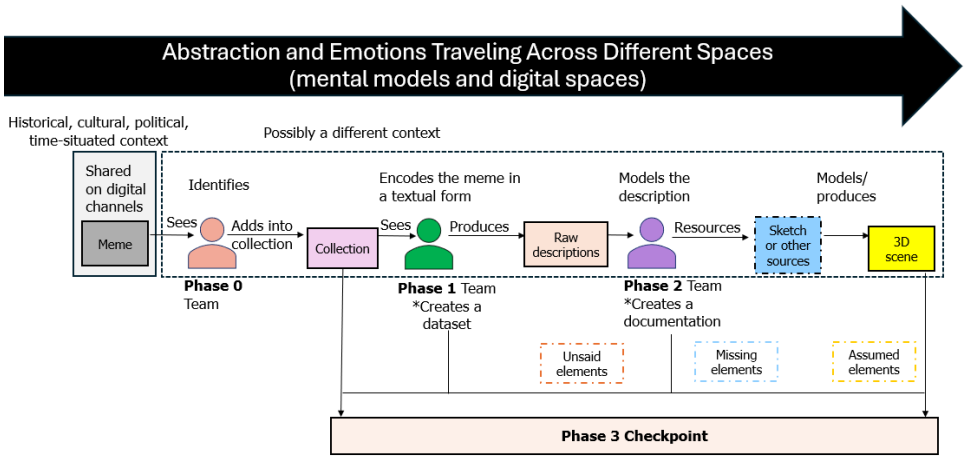


Fig. 3. An illustration of how abstraction and emotions travel different spaces within the project Phases 0-3. *Unsaid*, *missing*, and *assumed elements* are identified within dashed boxes.

Phase 1 includes creating the Observer-Centered Dataset attributes (see Section IV-C), which are split into three dimensions: 1. Concrete Design, 2. Blend, and 3. Emotional Design. The Concrete Design dimension covers objective elements that facilitate an image identification; the Blend focuses on the image's observer, whereas the Emotional Design on an image's messy layers of emotions (see Section VI). As we place the observer as an image's target, the Blend dimension blends together the three dimensions, see Figure 9. Finally, our dataset's **focus on *the observer* and *networked emotions* is its most distinguishing feature** – we will analyze our dataset in future work.

We noticed key linkages with assistive technologies during our investigation of image description methods. Hence, our work is inspired by the *Accessible Publishing* [32] advice on how to write accessible image descriptions for people with *print disability*, "which includes individuals who are blind or visually impaired, people with cognitive and comprehension disabilities, and persons who have physical mobility challenges" (see in Section II-C our parallel with alt-text).

In Figure 4, we illustrate our process for a human interacting with an image aiming to describe it: we depict the process as a ladder, starting from the initial viewing of the image and collecting information as we move up to finally reach a total understanding at the top. We follow a narrative approach to describing the images as we combine the guidance from three resources, all explained below:

1) Methods and advice from [32].
2) Heuristics from [33] to capture the whole image's meaning in an image description.
3) Advice from [34] on how to describe memes.

Our preliminary image descriptions' version was similar to the *Accessible Publishing's* [32] long descriptions, which are detailed textual descriptions that can be "several paragraphs long and/or may contain other elements such as Tables and lists. This technique is generally used for complex images where spatial information needs to be conveyed to the reader such as maps, graphs, and diagrams. Sometimes called extended description, these descriptions are too long and complex for alt-text."

However, as our process evolved, we felt the need to create two kinds of descriptions: raw and detailed – giving rise to the *unsaid elements*, which are the elements to be removed/filtered out of a detailed description to create a raw description. In other words, the elements to be "unsaid". Finally, instead of describing complex maps/diagrams, our detailed descriptions aim to describe images that rely on abstract reasoning mechanisms characteristic of memes.

Nganji and colleagues [33] propose heuristics to capture the whole meaning and description of the image. It addresses the "who", "what", "when", "where", and "how" of the image: "*who* asks the questions relating to the people in the image, while *what* relates to other non-human objects including buildings, trees, automobile, etc. including their descriptions such as colour. *When* on the other hand asks questions related to time such as when the picture was taken (time, season, etc.) while *where* seeks to find out the location such as where the image was taken, the positions of various objects in the image, etc. *How* relates to actions, emotions, etc". The authors propose an Image Description Assessment Tool, which is a Java-based tool for assessing how well an image description matches the actual content of the image on the web (it also provides a speech interface so that people can listen to the description of an uploaded image); thus, weights are applied to the heuristics categories to determine how close a description is to the original image.

To conclude, we are inspired by Lewis [34] specific strategies for describing memes, summarized as follows: 1. write any text that precedes the image; 2. describe the subject briefly (who or what is depicted) and 3. note any alterations in the subject's appearance, if relevant. 4. explain what the subject is doing; 5. Be explicit about the punchline.

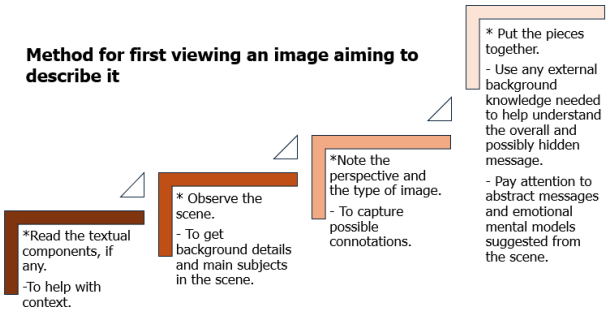


Fig. 4. We depict, as a ladder, our process for interacting with an image aiming to fully describe it. The process starts from the lower step and finishes at the top.

B. Detailed Textual Descriptions

Our detailed descriptions are created according to the instructions below (it would be interesting to run human studies and investigate if these instructions can assist in creating accessible image description tools):

General Instructions. Prioritize describing an image simultaneously with viewing it for the first time to ensure a fresh perspective and avoid leaving details out. Write a description as a "building up" process, ultimately leading to an overall understanding of the image, as Figure 4 shows. Be as detailed and precise as possible with your visual and emotional explanations, but be sensitive to cultural differences. There is no need to describe things we assume to be common knowledge (as long as the visual details match those), such as the shape of objects (or even color, as in the case of a polar bear, which frequently *appears* to be white). Then, the specific steps are:

1) Provide a general overview of the type of image and the main subjects, e.g., "This image is a (photograph/drawing/etc.) that depicts a (cat/person/plate of food/etc.)"
2) If present, describe the location of any text in reference to the image, write the text verbatim, and note the original language. E.g., (above/below/etc.) in the image, there is text (originally in English/Portuguese/etc.) that reads "...".
3) Provide the subjects' and the scene's detailed description. E.g., position, actions being performed, colors, materials, etc. Note: some details are left out, such as the color of the ocean, as long as details match what is assumed to be common knowledge.
4) Optional: if the image's perspective is necessary for overall understanding, include that. E.g., in the case of food on a plate, a top view is most likely essential to see all the elements clearly. So, one would describe, "The perspective of the image is directly above the plate of food." – We initially assumed the image's perspective was unnecessary; however, while it may not be the most critical aspect of interpreting an image, it may support connotative meanings in memes.
5) Connect everything together, explaining the punchline. Important: provide the context and set up the scene first to allow readers to discover the punchline by themselves before reading this part of the description. Explain the humor/emotional elements/meaning of the image and provide any additional details/context/pop culture knowledge/background information if necessary. E.g., "There is an urban legend in Brazil about a ghost who haunts bathrooms, and there was the pandemic shutdown. Thus, the joke is that the ghost is upset by the absence of students in the school's bathroom to scare." Without these two pieces of context, the corresponding meme does not make sense. This step is the most open-ended and complicated, as you need to analyze the emotional layers in an image, as well as any crucial outside information; also, it varies based on the image observer and interpretation – one person might find an image funny, and another might not.

To conclude, detailed image descriptions must both a) **Accurately** capture the image's explicit and concrete elements, and b) **Reliably** capture the image's abstraction, coherence, and contextual bridges needed to convey its meaning. Then, as a message travels through different spaces (see the box below), it should remain **consistent** with the image's concrete elements and abstraction:

> Meme → Phase 1 teams' mental models → detailed description → reader

On the other hand, raw image descriptions must a) Encode the bare minimum amount of the meme's explicit and concrete elements in a way that only just allows a reader to get the message's meaning. b) Lack of any reference to how abstraction, emotions, and connotation are used to build a message. The purpose is to check how well one recreates the message's overall meaning from only raw pieces of it. Therefore, checking how abstraction and emotions travel different spaces, as illustrated in Figure 3.

Hence, a message should remain **consistent** in spite of traveling through spaces (see the box below). We hypothesize this will help us investigate how humans get connotative meanings, abstraction, and emotional messages from missing information.

> Meme → Phase 1 teams' mental models → raw description → 3D teams' mental models → 3D scene

C. Raw Descriptions and the "Unsaid Elements"

A raw description is created by "breaking" or "tearing down" a detailed description through filtering out what we call by *unsaid elements*. One could think of using **alt-text**, as there are "no hard rules on how long alt-text should be, but they are usually a short phrase or at the most, a couple of sentences" [32]. Although that is somewhat similar to raw descriptions, as they are short image descriptions, they are not the same: raw descriptions are more vague and lean than alt-

text since they are used to check if/how a reader (the 3D team) catches and portrays the *unsaid elements*.

Thus, elements such as colors, shapes, descriptions of what things look like, affordances [35], [36], and explanations of the emotional/humorous components are all removed. Following [7] take on [36], we define an affordance as a "relation between an agent's abilities and the physical states of its environment" [7]. As we improve our methods, we will add the following step: **check/rewrite the raw description to make sure the used words have a clear denotation and as few as possible connotations**.

The instructions below detail how to write a raw description – note that we are moving down from the top of the ladder depicted in Figure 4:

1) Remove any explanation of humor/emotional elements/image meaning, as well as any explanation of background information/context (*i.e.*, exclude everything written for the detailed description's step 5).
2) Determine whether to remove perspective, if present. E.g., there is an image of a snake in a banana peel, and since perspective is unnecessary for interpretation, it should be removed. However, it should be kept otherwise; e.g., in another image, a fancier car must appear in the front to seem like it is the image's focus and "trick" the observer; another example: there is an image of eggs on a plate, in which perspective helps to understand that the eggs are supposed to look like two people holding hands. Still, perspective is somewhat open to interpretation, and justification should be documented for the choice, whether perspective is included or not.
3) Remove any visual details that do not interfere with understanding the image, such as color, type of material, and any non-objective descriptions. Frequently, the shape or color of an object, such as a chair or a Table, does not interfere with the image's interpretation. Therefore, those only remain if absolutely necessary to convey the image's meaning; e.g., in an image, the orange color of a cat's fur provides visual cues for it to look similar to a croissant, and the same comparison would not be as evident with a different color – for example, see the "croissant-cat image" in [37].
   - Remove mentions of the number of objects (as long as they are not necessary to convey the image's meaning). In the "machine learning" meme (see Section IV), the specific number of computers in the classroom is considered an *unsaid element* and therefore replaced with "rows of computers". In this case, we are checking whether the 3D modeler understood that the computers are meant to imitate students in a classroom. Also, details about the specific type of computers are removed, as most computers would still convey the appropriate message.
   - Remove/replace unnecessary details about image subjects. E.g., an image shows a little girl wearing glasses; however, since that is not needed to convey the message, we used the word "child" instead.
4) Text location and wording are objective and should thus remain as-is in the raw description.
5) An image's type (e.g., photograph, drawing) is *almost* always removed since the image's recreation is 3D modeled. However, there are exceptions, such as when the medium helps to drive the image's meaning. For example, if an image's element was clearly drawn with simple black and white lines, and such detail is needed for understanding, it should be kept in the raw description.
6) Finally, clean up and observe the used language to not give details away. For instance, in the eggs on a plate image, we made sure to keep the raw description as objective as possible. Instead of writing the yolk has a "smiley face", we wrote that there are "two dots next to each other, with an upward curved line underneath". It is vital to identify and filter out these "micro interpretations" and leave it up to the 3D modeling team to realize that the yolk has a face. Make sure to avoid repetitions: analyze the raw description and remove any general introductory descriptions if they repeat information unnecessarily.

"When are you done writing a raw description? What determines that it is finalized and ready to be sent to the 3D modeling team?" That is not a trivial question, as sometimes the team felt the need to keep cleaning up raw descriptions within multiple iterations. The team engages with explicit and implicit knowledge as they identify the *unsaid elements* – Zheng *e*t al. [38] summarize [39]: when "knowledge has been articulated, then it is explicit knowledge. Otherwise, another question is raised: Can it be articulated? If the answer is yes, then it is implicit knowledge. If the answer is no, then it is tacit knowledge".

Still, we consider the task to be complete once a raw description does not include any unnecessary elements. Therefore, it basically has the subjects of the image, any text if present, and the bare minimum for other details. It does not over-describe visual elements, does not hint at the image's meaning/humor, and does not emphasize an element as more important than the others.

Preparation for Phase 3, checkpoint. In parallel to identifying the *unsaid elements*, we document a "checklist" of things to look for in the 3D scenes and check if/how the 3D modelers depicted the messages' overall meaning. Once we finish processing the Phase 0 images, we will have a collection of *unsaid elements*, which we hypothesize will help to create a visual Dictionary that refers back to emotions, abstraction, and connotative meanings (Phase 5).

Finally, we will process all memes within our collection but model 3D scenes from a subset only. Then, in Phase 3, we will check how the 3D team filled in the gaps from missing information and interpreted both the meme's main visual components (concrete elements) and the abstract and emotional components. Two potential lines of inquiry for Phase 3 are as below:

1) Llorens-Gámez *e*t al. [40] show that components, such

as form and geometry, space distribution and context, color and texture, among others, influence memory and/or attention, and can be assessed objectively. The verbal description of a sink may bring up very different mental models based on each individual's background, as architecture differs across countries and cultures. It would be interesting to investigate to what extent familiar shapes or contexts populate a 3D modeler's *assumed elements*. If a modeler is used to seeing wood-made and square-like sinks, will those occupy the *assumed elements*? (Of course, other players are in place, such as how easy it is to design that shape and texture in the chosen 3D modeling tool.)

2) Leshin et al. [41] provide preliminary evidence that brain representations of emotional facial expressions are influenced by two sources of conceptual knowledge: a person's access to emotion category words and their cultural background. Their findings support evidence that conceptual knowledge activated in the minds of perceivers influences emotion perception. If an emotional context is related to disgust in the modeler's culture but anger in the original image's culture, will the 3D scene still be consistent with the original image? Images that are meant to be humorous to some may not be to others because humor shifts in different cultural contexts – see [42] for a view on how cultures create emotions or [43] for findings suggesting that emotion depends on context, culture, and their interaction.

III. Methods: 3D Scenes from Raw Descriptions

In this Section, we describe our processes for modeling a 3D scene from raw descriptions. For clarity purposes, in this Section only, we use interchangeably 'description' and 'raw description.'

We examined several 3D modeling tools before selecting Blender Version 3.5.1 for its flexibility and learning curve – see [44] for a review on Blender's versions and interfaces, and [45] for an application built on top of Blender. As stated earlier, modelers do not have access to the images that originated descriptions. To prevent influencing each other's style and approach, they individually worked on the 3D scenes. Finally, the glossary terms (Appendix A) are key to interpreting the Gateways diagram, shown in Figure 6.

We ask questions such as: "How to model an AI system that *gets* abstract and emotional messages from spatial communication? What does that even mean?" Hence, memes are a key resource, given their use of spatial communication to convey a message. According to Tversky [46], by using position, form, and movement in space, gestures, and actions convey meanings. In that sense, differently from solely symbolic words, visual communication can directly convey content and structure (both literally and metaphorically). Although it may lack the rigorous definitions words can offer, visual communication delivers flexibility and suggestions for meanings. Such flexibility, in turn, requires context and experience to interpret conveyed meanings [46]. At the same time, "the layout of the physical environment, including the apparent steepness of a hill and the distance to the ground from a balcony can both be affected by emotional states" [47].

Cohen and colleagues [48] detail four technological affordances represented in research on emotion: interactivity, personalization, accessibility, visibility, and social cues; finally, the authors discuss how technological affordances relate to emotional regulation via media use. *Social cues* are particularly important in our project since they are nonverbal signals that "infuse meaning into messages, including information about a sender's emotional state"; and "Technologies vary in terms of the type and number of the social cues they afford to users for emotional expression" [48]. That context helps to answer a question such as below.

Why the use of 3D scenes? We chose a 3D format since it provides different perspectives and enables people to play with the objects on the screen, enabling a richer experience (this interactivity is unique to 3D spaces compared to 2D images, while there are mixed results on its advantage for learning [49]). In addition, we can take screenshots of a 3D scene if needed (as in Figure 8). Finally, we sought to investigate how the modelers translate a raw description into a dynamic encoding (dealing with spatial organization and hierarchy), which opens avenues for applications within spatial thinking skills.

Instructions to create a 3D Scene are shown below:

1) Read the description and create a 3D scene as close as possible to your comprehension of the description.
2) You are free to sketch your ideas and to search online for reference images if that helps the modeling process (e.g., to model an airplane or some other unfamiliar object).
3) Do not search for memes and do not observe the other modelers working on their models – so that you do not influence each other's style.
4) Focus on developing your own shapes and avoid, as much as possible, importing shapes and libraries into the 3D modeling tool.
5) Engage with your peers to share tips on the modeling tool.
6) Finally, focus on portraying what you interpreted the message to be. Do not focus (or spend your time) on creating fancy-looking 3D scenes.

Often, modelers felt the need to use sketches either to make sense of the description, fill in the gaps, and/or visualize objects' details in different dimensions and perspectives (more in Section III-A). In that case, the description and sketch are revisited during the modeling process.

"How do you decide that a 3D scene is complete?" We investigated that question and concluded modelers were following two main gateways to evaluate and decide if a 3D scene was complete. We named those "accuracy" and "faithfulness" gateways, see Section III-B. Our focus relies on the description's message but not on creating fancy-looking 3D scenes. Hence, striking a balance between time and detail in the scenes was crucial. Also, evaluating each model's *accuracy* and *faithfulness* helped determine when the modeling process was complete.

Accuracy is assessed by revisiting the description (and sketch if used) to check if the description's explicitly stated elements have been included in the model. *Faithfulness* is assessed more complexly by evaluating the emotions in the final scene and checking if these align with the interpreted emotions from the description.

Since cultural background and experiences play a key role in how individuals make sense of a description and encode it into a 3D scene, we briefly provide our cultural context. Our research is being conducted within the USA Liberal Arts institution's cultural context, and we are a multicultural team: in addition to the US, some people lived or are originally from countries such as Brazil, China, Japan, and Thailand, and so far, six modelers (undergraduate students) have worked on this project.

A. Sketches

A sketch is a drawing draft that helps the modeler brainstorm and navigate through mental models triggered by the description, usually made before building the actual 3D scene. A sketch is an external representation [50], a visual-spatial display that augments cognition [51].

"When people read text, they construct representations of several levels, including" text-based representation (a representation of the text itself, the propositional content of the text) and "a representation of the situation or object described in the text" [52]. Depending on different descriptions, the modeler's mental models may be easily formed as a whole scene, or they may initially appear as separate objects or parts and need to be joined together after considering their relationships with each other – all these considerations are recorded in the modelers' written documentations. Modelers also reflect on any elements that feel to be missing from the description for the scene to make sense. Hence, modelers can combine potential *missing elements* into their mental models.

In addition, they may need to search for certain objects or parts to draw or model details successfully. Reference images help to visualize objects that are unfamiliar or hard to imagine (e.g., an armadillo body). Then, they can draw the sketches according to their mental models and the reference. The drawing process may be done with pen and paper or with digital drawing apps. The sketch will usually be a simple line drawing with black drawings and a white background.

Modelers review the description and check with their sketches, and they may continue to identify what is missing in their sketches and adjust it accordingly. They can also add annotations to help them better model their drawing and document their decisions. The primary purpose of having a sketch is to facilitate the process of building a 3D scene. Specifically, a sketch can help the modelers in three ways:

1) **Augments Cognition.** As the modelers navigate through mental models triggered by the description, the initial product can be vague and blurry at first glance. However, the process of drawing can help to consolidate ideas and make them clearer. Modelers can externally visualize the scene they are considering, enabling them to review their mental models. It also helps them think about what is missing from the description and elaborate on this.
2) **Reference.** It provides a standard reference for the 3D scene that boosts the modeling process efficiency. A modeler may find that transitioning directly from mental models into the actual 3D model can be difficult, so a sketch acts as a bridge. For instance, Blender allows modelers to import reference images; thus, they can import sketches into the tool and build the 3D scenes according to the sketch. A sketch also provides a standard for the size of the objects, the proportion and layout of the whole scene, and how the model parts relate.
3) **Consistency.** It may be easier to maintain consistency between mental models and the 3D model if there is a sketch to compare with. The 3D scenes must be consistent with the modeler's initial mental models of the scene. Thus, the 3D scene reflects the modeler's interpretation. However, many factors may decrease this consistency, such as technical issues with the modeling tool and the difficulty of different models. In this case, a sketch helps to record a modeler's initial mental models after reading the description, as the drawing process tends to be more flexible than 3D modeling.

In Figure 5, we show a sketch with a scene's different perspectives. Before drawing it, the modeler read the description of a polar bear standing on an iceberg in the middle of the ocean and formed mental models of this scene. The description mentions a reflection of the polar bear's skeleton on the ocean's surface, so the modeler considered the different perspectives and what should be seen under each perspective according to the physical properties. The modeler searched for images of polar bears and icebergs to observe details to draw them better and draw the scene by combining mental models with details from reference, real images. First, the modeler drew the scene from the front view, where the reflection of the skeleton cannot be seen. Then, from the top front view, the skeleton can be seen on the surface of the ocean. Additionally, there is a sketch of the skeleton itself to show its details – in Figure 8, we show the modeler's corresponding 3D scene (the image that inspired the corresponding raw description is shown in Figure 7).

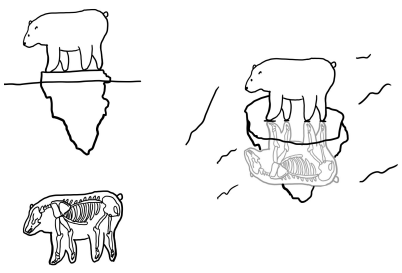


Fig. 5. An example of a sketch that contains a scene's different perspectives. The modeler sketched those to inform the 3D modeling process.

B. 3D Modeling Outcomes and the Gateways Diagram

We first detail the Gateways diagram and then show, in Section IV, our 3D scenes' static images. Modelers engage with networked emotions and emotional mental models as they switch between and across mental models to guide a description's sense-making and decision-making that leads to creating a 3D scene. Although we understand that "cognitive and emotional mental models are activated at the same time" [9], we bridge the modeling task with a dual-process account of decision-making [53], and each process has its own gateway. We decided to do so to account for both the way modelers' described their processes and highlight the importance of emotional mental models. Similarly, modelers transit between explicit, implicit, and tacit knowledge in both gateways.

Inspired by work on diagrams and cognition such as [53], [54], and [55], we designed the Gateways diagram (Figure 6) to understand the modelers' decision-making process and how they navigated the 'layers' or dimensions of emotional processing during the 3D modeling process. In the end, our diagram was informative not only in understanding the modeling process but also in checking for consistency across modelers – we will add the diagram to the 3D modeling instructions in future work.

As the diagram shows, modelers start by reading a description, and their purpose is to pass the *accuracy* and *faithfulness* gateways to complete a 3D model, producing a 3D scene. Generally, the beginning process (diagram's top/first half) tends to focus on *accuracy*, and *faithfulness* is prioritized towards the end of the modeling process (diagram's second half). We designed the diagram to allow for flexibility in the 3D modeling process, as modelers seek to create accurate and faithful 3D scenes that capture both visual and abstract details. The gateways are not completely divisible, and the process of addressing each gateway is open-ended. Therefore, achieving accurate and faithful 3D scenes can look different for distinct modelers or even for the same modeler on different days. However, despite the open-endedness, a scene must be consistent with the description.

Frequently, modelers started from the explicit and concrete elements and launched a Raw 3D model. By this point, models can be checked for the *accuracy* gateway. They may search for clues, such as image references, or sketch a few ideas to help identify *missing elements*. Modelers decide whether the model sufficiently reflects the observable visual features of the mental model they created based on the description. By 'passing' the *accuracy* gateway, they ensure that the model is accurate with the description.

At some point, modelers make assumptions to turn *missing elements* into *assumed elements* they can incorporate into the model (diagram's 2nd half). Likely starting from implicit knowledge to identify *missing elements* but then engaging more with explicit knowledge to instantiate *assumed elements* and incorporate them into the 3D model. Modelers use imagination (see [56] for a detailed view of human imagination), mental models, and knowledge/experiences to make assumptions, turning *missing elements* into *assumed elements* that can be added to the Raw 3D model to complete the model. They may sketch to reflect on different facial expressions or search online for a clue (for example, to investigate: "how does a happy turtle look like?"). Finally, the modeler documents assumptions that guided the specific *assumed elements* and other notable details about their decisions throughout the modeling process – we will investigate that documentation in phase 3.

As the model nears completion, modelers frequently focus more on emotional mental models and networked emotions. That helps examine the *faithfulness* gateway: the modeler reflects whether the 3D model sufficiently reflects the description's abstraction and emotional tone. Once the modeler decides that the model passes both gateways, the modeling process is complete, producing the 3D scene: a completed 3D model that is accurate and faithful to the description and to the modeler's mental models resulting from the description. In Section IV we show a sample of our 3D scenes and sketches.

In [1], we illustrate possible questions modelers may ask themselves while modeling emotions. The questions refer to 'layers' or dimensions of emotional processing during the 3D modeling process: modeler's, 3D model's encoding, observer/audience, and image's *via* raw descriptions. Once an observer views and interacts with a 3D scene, if the observer's overall response matches the *Observer's Intended Emotional Response* (see below), abstraction likely made it successfully through spaces and gateways. It would be interesting, in future work, to run human studies in that direction.

Modelers reported that some descriptions were harder to navigate since they evoked multiple emotional mental models to make sense of - particularly when there were conflicting or unaligned (*messy*) emotional layers. Given that challenge, we list below some of the guiding questions that helped ground the modeling process:

1) **Source.** From the raw description, what can I assume about the image?
2) **Mediated Communication.** Given my experience with popular culture, social media, and memes, what does it seem to mean? Is this supposed to be humorous? How do I feel about that?
3) **Characters.** Are there multiple characters? Do they have aligned, neutral, or unaligned emotions? Who is the main subject?
4) **Modeler's Emotional Response.** Emotions triggered in the modeler as part of making sense of the description and finalizing the 3D scene.
5) **Observer Intended Emotional Response.** Emotions the observers are supposed to have when viewing the 3D scene for the first time. How is an observer supposed to feel? (Should that be similar to how I felt when I read the description?) Should that be aligned with the characters in the scene?
6) **Emotional Mental Models.** Given the raw description and *assumed elements*, it is time to put on multiple "emotional

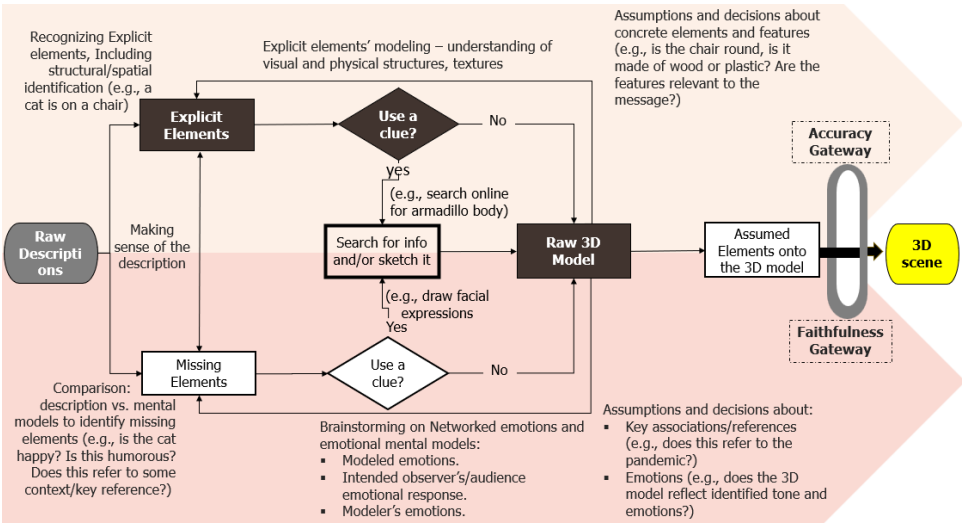


Fig. 6. Gateways diagram - Process to create and decide that a 3D scene is complete: it has to pass the Accuracy and Faithfulness gateways (right).

hats", deal with the messy layers, and model the scene and its components.

IV. A Glance at the Project's Outcomes

We show three examples of our detailed textual descriptions along with *unsaid elements*. In Figure 8, we depict the corresponding 3D scenes' static images and raw descriptions in Table I. We named the examples for reading purposes – the Phase 2 team receives ID numbers only. Finally, we present the Observer-Centered Dataset in Section IV-C.

A. Raw, Detailed Descriptions, and 3D Scenes

As the three examples below show, the identification of *unsaid elements* includes a reflection of what needs to be removed or reframed to create a raw description. We consider that reflection insightful for identifying clues people may use (potentially without realizing) to comprehend a message and ultimately key in reflecting on human-AI interaction.

Finally, as mentioned in Section II-C, we will add another step in our process to ensure that raw descriptions' words have a clear denotation and as few as possible connotations.

Example 1: "Machine Learning", ID 13.

A "machine learning" meme can be found, for example, in a Reddit post [57].

Detailed Description. This image is a photograph that depicts a white room with upright computer screens (no keyboards). There are three rows of computers, with three computers per row, on a wooden floor facing the front of the room. In the front of the room, there is a larger screen facing the rows of computers. On the large computer screen, it shows text that says "machine learning" implying the joke that the computers are learning by being in a classroom setting like humans. The perspective of the image is low to the ground, behind, and to the left of the rows of computers.

Reflection: Unsaid Elements. How many computers are needed to convey the sense of a classroom? What types of computers come to mind? Elements: room's color, floor, type of computers, number of computers, the fact that there are no keyboards and an explanation of the joke/background knowledge. To understand the joke, one would need to know what a standard classroom setup looks like, and a basic understanding of what machine learning is. **Is perspective important?** Yes, to ground the metaphor.

Example 2: "Polar bear", ID 18.

The "Future we all face" cartoon, or "Polar bear" meme [58], see Figure 5.

Detailed Description. This image is a drawing of a polar bear balancing on a small iceberg in the ocean, with the sky as the background. The image conveys a sad message. The polar bear's four feet can barely fit on the iceberg. Its bottom is pointing towards the top right of the image, and its nose touches the water. The water shows a reflection of the polar bear on the iceberg, but as a skeleton. We believe this image is intended to provide dark commentary on the state of global warming and the polar ice cap melting; we think the reflection is meant to be a window into the future extinction of the polar bear population and perhaps ours. **Is perspective important?** Yes, because both the bear and reflection must be seen to understand the message.

Reflection: Unsaid Elements. Orientation of the bear's body, shape, and color; the same goes for the sky, ocean, and what a reflection is. Climate change understanding and how it relates to a polar bear, the connection between body and skeleton. The bear is "facing" a skeletal version of itself, highlighting a possible reality, which conveys another layer to the image's connotations.

Example 3: "Butterfly", ID 22.

This description is based on a pun that generates the word "butterfly", see [59].

Detailed Description. This image is a photograph of illustrated elements and real insects on a paper. On the left of

a paper, there is a simple black line drawing of a person's behind, starting at the waist and ending at the top middle of the thighs. In the middle, the letters "ER" are written. To the right of those letters, there are two flies placed on the paper. The perspective of the image is directly above the paper. The joke is that it is intended to represent the word "butterflies". **Is perspective important?** Yes, because the image includes drawings that can't be seen from a side view of the paper.

Reflection: Unsaid Elements. The number of flies (how many flies are needed to convey the message?) and their appearance. How the objects in the scene visually spell out the word "butterflies" and a person's behind creates the beginning of the word. How different elements merge/blend to create a single word.

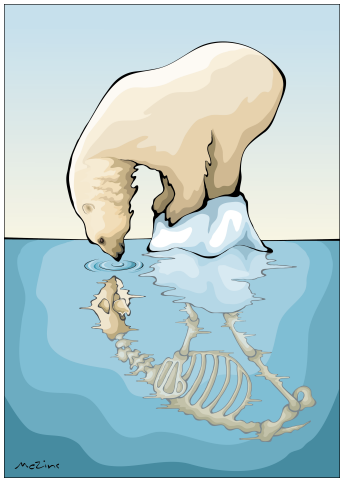


Fig. 7. The future we all face, by Mary Zins [58] (used with permission).

B. A sample of 3D scenes along with corresponding raw descriptions

In Figure 8, we show a sample of our 3D scenes, sketches and corresponding raw descriptions in Table I. Although the modeling task may seem simple, modelers faced insightful challenges along the way. For example, some descriptions mention cultural references the modelers did not recognize, and another interesting barrier came from a description centered around veganism, with a sarcastic tone. Since veganism was not a common reference to the modeler, it was challenging to understand the atmosphere the description created. Therefore, in addition to cultural background, contextual knowledge (e.g., context brought by COVID-19), popular culture, and the media were often needed to make sense of the descriptions – which was expected, given our focus on memes.

Some descriptions contained references to popular movies that must be understood for the description to make sense – there is a description that combines the context of the pandemic with the 2000 movie Castaway *via* the ball the main character bonds with. Background knowledge of the movie not only informed the modeler of what the face should look like but also what it meant in the context of the text (see Figure 8, lower left).

When we launched the project's phase 1, we were still examining what a "raw description" should look like. As we experimented with continuously removing details, we finally decided on a final method. However, we decided to keep older versions to record our trajectory. For instance, notice Table's I first row, which is derived from an older method of creating raw descriptions. To conclude, an observation on the memes that inspired #12 and #20: a) We were unable to retrieve an online source for the Brazilian version of #12. Thus, we provide an English version found within other COVID memes [60]; and b) Multiple versions of #20 are described in [61] and [62].

C. The Observer-Centered Dataset

Phase 1's goal includes the identification of attributes to examine many memes at once. Here, we present our dataset dimensions (see the dataset attributes in Appendix B). We conducted a data-driven approach to identify ad-hoc categories and image attributes, similarly to [63], whose work provides methodological directions for the study of memes.

As we kept cataloging new attributes and writing descriptions, we saw the need to better organize them, leading us to group the categories within three dimensions. Therefore, each dimension covers a set of categories, and each category has a set of attributes. That organization assisted us in capturing the images' observer experience and the interplay between concrete and abstract elements, and networked emotions.

Our approach to creating the dataset is similar to [63], which asks two questions: "Which meme formats are currently circulating online?" and "How do popular meme formats convey their message?" to then propose a methodological toolkit to analyze Internet memes. Giorgi [63] conducts a data-driven approach to create eight ad hoc categories to examine a sample from a dataset of static images collected on Instagram within the Italian cultural context. Although similar, our work presents important distinctions (in addition to the languages explored), such as our focus on abstraction and emotions, leading us to consider the observer's experience.

Similarly, Cochrane et al. [64] create a dual classification system for meme categorization: meme composition and multimodal quality. Meme composition focuses on a meme's structure, i.e., on how memes recontextualize images and text to create new meanings, whereas multimodal quality on the ways that text interacts with the image. Although the authors also consider an image's structure to get into a meme's meaning, our approach is different, given our focus on how a message travels through spaces.

Currently, the Observer-Centered Dataset has 26 attributes distributed within three dimensions. The dimensions are: Concrete Design, Blend, and Emotional Design. The Blend dimension centers around an image's observer, while the Emotional design on networked emotions. These two dimensions share two categories (Emotional Alignment and Humorous Intent), as Figure 9 shows, and cover categories that are human-interpreted and more flexible than the Concrete Design dimension (shown at the top of the Figure).

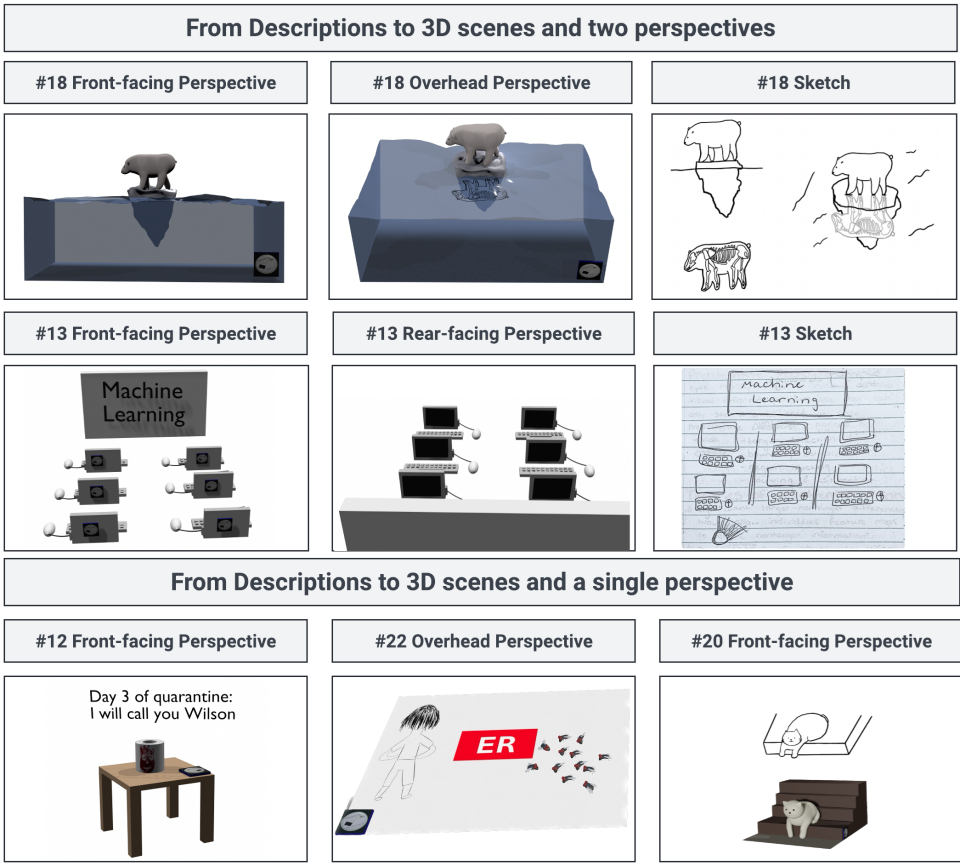


Fig. 8. Phase 2 and a sample of 3D scenes' and sketches' from a raw description. Numbers refer to the corresponding raw description ID.

Process and attribute labels.

As we added memes' descriptions and identified new attributes, we faced challenges in defining concrete labels and definitions. It was difficult to create a dataset that covered the meaning of any meme-like image (within our collection scope) without excluding important details of some images or including attributes that are irrelevant to others. We kept adding new attributes as additional images were processed and reshaping older attributes, but the attributes were not always relevant for all of the images, and some were too ambiguous.

For instance, it was difficult to name the image attributes in a concise way that reflects their meaning. This challenge is explored in [65], which presents the idea that words have different meanings depending on the individual. Their results show that "at least ten to thirty quantifiably different variants of word meanings exist for even common nouns. Further, people are unaware of this variation, and exhibit a strong bias to erroneously believe that other people share their semantics. This highlights conceptual factors that likely interfere with productive political and social discourse". Their findings support our hunch that categorizing abstraction and emotions using attributes containing one or two words is challenging, as different interpretations of words can hinder understanding, especially in the context of dimensions meant to convey abstract/interpretive attributes.

A note on irony. Lozano-Palacio et al. [66] provide a broad cognitive-pragmatic perspective on the irony that interprets "ironic meaning" as a result of complex inferential activity that arises from conflicting conceptual scenarios. They distinguish basic and re-adapted uses of irony; basic uses are: Socratic irony, rhetorical irony, satirical irony, tragic irony, dramatic irony, and metafictional irony. Irony is then "determined by the attitudinal element arising from the clash between an epistemic and an observable scenario". We follow the authors' approach and consider verbal and situational irony as different materialization of *the same phenomenon*: "In both cases, the epistemic scenario is drawn from the speaker's certainty about a state of affairs (be it formed through an echo or not), and the observable scenario from the situation that is evident to the speaker" [66].

Especially if focusing on humor, the layers of emotion do not always align with the image observer and the characters in the scene. For example, an image was clearly conveyed through the detailed description "Photograph with text at the top stating 'My cat isn't paying enough attention so I improvised.' We see the back of an orange/brown cat's head with its ears up and half of its body facing away from the camera. The cat's head is to the left of the image, and its body is to the right of the image. It appears to be sitting on a couch, with the background showing part of a door and some

TABLE I
RAW DESCRIPTIONS PROVIDED TO THE 3D MODELING TEAM CREATE THE 3D SCENES DEPICTED IN FIGURE 8.

ID	Raw Description
12	An image of a roll of toilet paper standing on one end, with a drawing of a red handprint with a face oriented vertically along the roll's position. There is leading text that reads: "Day 3 of quarantine: I will call you Wilson". The image references the external context of the movie Cast Away (2000) and a volleyball which is given a handprint and face and is then named Wilson by the character when they are isolated on an island. The perspective is forwards toward the toilet paper roll, which sits on a paper towel laid out on a table.
13	There are rows of computers on the floor facing the front of the room. In the front of the room, there is a larger screen facing the rows of computers. On the large computer screen, it shows text that says, "machine learning". The perspective of the image is low to the ground, behind and to the left of the rows of computers.
18	A polar bear balancing on a small iceberg in the ocean. The polar bear's four feet can barely fit on the iceberg. Its nose touches the water. The water shows a reflection of the polar bear on the iceberg, but as a skeleton.
20	There are two sections, one above the other. On the bottom section, there is a cat with most of its body on a large step. Its hind legs are under the body, and are not visible, and its tail is also not visible. Its front legs extend directly down from the step, resting on the floor. The perspective is slightly above and slightly to the right of the cat. On the top, there is a simple line drawing imitating the shape of the cat from the bottom image.
22	On the left of a paper, there is a sketch of a person's behind. In the middle, the letters "ER" are written. To the right of those letters, there are real flies placed on the paper. The perspective of the image is directly above the paper.

shelves. On the back of the cat's head, there are two googly eyes facing the camera. The joke is that the human had to put googly eyes on the cat to pretend that the cat was looking at/paying attention to them."

The image can also be successfully conveyed through the raw description: "Text at the top stating 'My cat isn't paying enough attention, so I improvised.' We see the back of a cat's head. The cat appears to be sitting on a couch. There are two googly eyes placed on the back of the cat's head." When it came to the dataset coverage of this image, interpretive challenges presented themselves, especially for humor alignment. To label the type of **Emotional-Alignment**, the emotion of the image's observer and the emotion of the image's subject must be determined, so they can be compared. But in the image, who is the subject? Is the subject the cat, or is it the human? This is a matter of opinion, so the image cannot easily/justifiably fit into the category of "aligned" or "unaligned"; therefore, we used the "ambiguous" data entry.

Also, **there are memes that call for an Outward (ad hoc) participant/observer**: they expand their scope as they incorporate us, outside observers, as if we were part of the image/meaning (as an illustration, consider the "Hand with Reflecting Sphere" by Maurits C. Escher). These kinds of memes informed us to center the Blend dimension around the image's observer. This dimension raises an interesting reflection: how to model an AI system that "sees itself" within a context and uses that to produce a holistic interpretation and successful predictive processing?

V. DISCUSSION

Project similarities with the Telephone Game come with caveats, such as the flexibility and open-endedness in modeling a 3D scene from a textual description. Still, that is exactly it: we seek to investigate how abstraction and emotions make their way through people (calling attention to a tension between *unsaid elements* and the audience's assumed elements) and foster ideas on developing *emotion-driven* AI systems and assistive technologies.

In Phase 3, we will compare the *unsaid elements* from phase 1 with the "missing" and "assumed elements" from phase

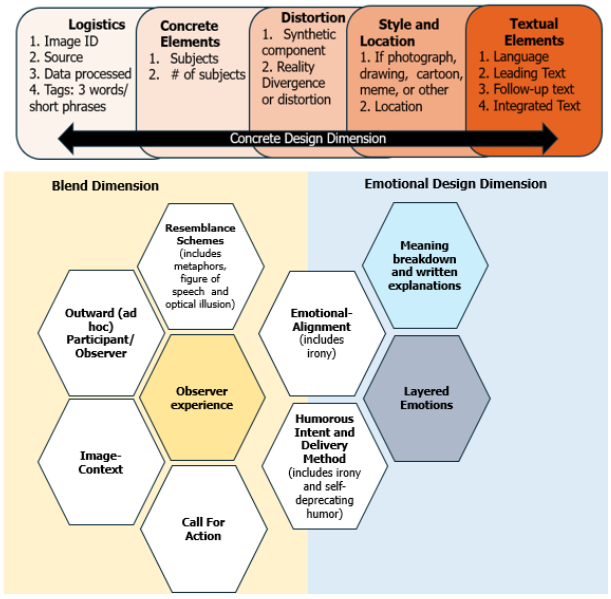


Fig. 9. The dataset attributes are categorized within three dimensions: Concrete Design, Blend, and Emotional Design. The categories *Emotional Alignment* and *Humorous Intent* belong to two dimensions: Blend and Emotional Design.

2. The amount to which they match will help us to reflect back into abstraction and emotions. Moreover, we are setting metrics to ensure objectivity, such as keeping consistent terminologies and processes across phases and building gateways; besides, concrete elements are easy to track across spaces (for example, check if a cat "made its way through spaces").

By translating an image's message into different presentation modalities (e.g., detailed and raw descriptions, 3D scenes), we are changing the conditions of comprehension [31]. Furthermore, with the removal of *unsaid elements*, a raw description becomes more abstract (it detaches as much as possible from the source image), making the message's comprehension task more abstract and open-ended (see the *concreteness effect*, in Section VI). "The advantages of concrete materials are that they can activate real-world knowledge during learning, induce physical or imagined action, enable learners to create

their own knowledge of abstract concepts, and activate brain regions associated with perceptual processing. The advantages of abstract materials are that they can focus attention on more useful functional features rather than superficial features and increase generalization across multiple contexts" [18].

Raw Descriptions and Local Coherence. As Schnotz [31] points out, texts are not carriers of meaning; instead, "they trigger processes whereby multiple coherent mental representations are constructed through an interplay between text-driven bottom-up and knowledge-driven top-down activation of cognitive schemata". Then, in a "text with only local coherence, successive sentences are semantically related but without an overarching thematic connection. In a locally and globally coherent text, successive sentences are connected and there is an overarching thematic connection." Detailed descriptions offer locally and globally coherent text, whereas **raw descriptions potentially offer local coherence but a weaker global coherence**, and readers "have to reconstruct the local and global text coherence in their minds" [31]. How raw can an image description be to still allow the image's message to be transmitted? Detailed descriptions make a task clearer: since the text is rich in details and coherence, the reader's task is *to comprehend* the abstract and emotional elements in context. On the other hand, although raw descriptions are shorter, they make a reader's task more abstract and open-ended, as a reader has to fill in the gaps using their background knowledge and experiences.

Generative AI. There is incredible work being done under the generative AI tools' umbrella, such as a foundation world model [67], adding voice and image capabilities [68], text-to-3D content creation [69], text-to-image [70], [71], and text-to-video [72], among many others. Our research substantially differs from those: we are not using artificial neural networks (or any other computational approach) to identify patterns and structures within data to generate content. Rather, our agents are humans, and we seek to investigate how a message's abstraction and emotional tone, among others, travel through spaces. Also, generative AI tools are not allowed in our project – except perhaps in later phases to compare our outcomes with a tool's output such as *genz 4 meme* [73], a tool that receives a meme as input, and outputs an explanation of inside jokes and hidden meanings (although, as of now, a tool such as GPT-4 has yet to develop robust abstraction abilities at human levels [74]). Likewise, although humans are our only agents and in Phase 2 we investigate what makes a modeler decide that a 3D scene is complete, our focus relies on the *message*, not on humans themselves - hence, for simplicity, we use 'mental models' as an umbrella term, and the Gateways diagram has a focus on the task rather than on the modeler.

Sentiment Analysis. Something distinctive about our work, in particular in relation to sentiment analysis (or opinion mining; see [75] for a review in computational sentiment analysis), is that we are not trying to determine if a message' emotional tone is positive, negative, or neutral. We are seeking to investigate if a message is kept consistent within spaces without necessarily classifying its emotional tone. In fact, we take into account 1) the messy layers of emotion [22], and 2) that humor shifts in different cultural contexts (e.g., images that are meant to be humorous to some may not be to others) – we acknowledge that cultures create emotions [42] and findings suggesting that emotion depends on context, culture, and their interaction [43].

Meme Sentiment Classification. While research is being done to classify hateful memes targeted at particular audiences [76], meme sentiment classification is still an area to be explored [77]. Perhaps our methods to break down images to write descriptions and our dataset attributes could be explored in that direction - for example, our dataset's focus on the observer could help to investigate if an observer is somehow being attacked through the message's tone or connotation (e.g., one could build on the *Outward participant/observer* attribute). Although we do not include typeface data in the Observer-Centered Dataset, **typeface effects are often used to convey strong connotation messages, and it would be interesting to investigate that further**.

Our dataset covers memes from Brazil and the US, imposing a unified view of both, which, although not ideal, allows us to compare them; we include information about the language of origin in the dataset as it provides additional context. When creating and reading the textual descriptions, we must consider the historical, cultural, political, and time-situated context of when an image was created (political memes offer a key example since they can become obsolete quickly). That potentially "interferes" with how a message makes its way through different spaces, and we are using the documentation created within project Phases to help us navigate that.

We launched this project with the aim of getting insights into the modeling of *emotion-driven* AI systems; still, our work offers applications for social impact and assistive technologies. Below, we provide a few ideas for further examination.

- **Learning from Meme templates → 3D Scenes.** We described the Gateways diagram, which serves as a guideline for the modelers to create 3D scenes out of missing information. What if we could combine meme templates with raw descriptions and our Gateways diagram's process to automate the generation of 3D scenes from memes? The 3D scenes could elaborate on the unsaid elements and have a focus on teaching a domain (e.g., computer science [3]) or assist with meme humor comprehension in adolescents with language disorder or hearing loss [78], emotion regulation in depression [79], spatial thinking skills, or help individuals with aphantasia [80] create various visualizations and explore them from various distinct.
- **3D Blueprints.** Inspired by [81]'s investigation of using text-to-image generators to create concept art for the 3D-modeling process of a character, it would be interesting to check if our processes to create a 3D scene can help to embed emotions and abstract concepts to design accessible interactive 3D blueprints for blind and low-vision people [82].
- **Humor Comprehension.** Similarly to [83], our work can inform the development of intervention resources to remedi-

ate humor comprehension deficit. In that direction, we would use as inspiration: a) Dr. Temple Grandin's strategies for creating concrete exemplifications of abstract concepts [84], [85], and b) Buxbaum et al [78], which moves from "old humor" (e.g., jokes, videos, and cartoons) to web-based humor (memes), and c) Dr. Spector's work on abstract language and cognition, which informed the creation of resources such as [86], [87] [88] and [89].

- **Descriptions and Cultural Sensitivity.** In this research, we propose a systematic approach to deconstructing memes into their fundamental elements and unsaid elements reflection. The breakdown helps identify an image's references/connotations and highlight key cultural and contextual knowledge, ultimately helping a description writer to 1) notice any gaps in the description and 2) ensure cultural sensitivity. Finally, as mentioned earlier, a key challenge in identifying the *unsaid elements* originates from making the implicit explicit, and we hope to inspire frameworks for identifying biases and microaggressions in visual content.
- **Diagnostic Images.** We wonder if the Phase 1 process of creating detailed, raw descriptions (in particular, the unsaid elements) and a dataset would help to inform the identification of diagnostic images [90].
- **Strategic Decision-Making and External Representations.** According to Csaszar and colleagues [91], there is work to be done to understand external representations' central role (visuals, more specifically) in the search for new strategies. Their research "highlights that the design and use of external representations — much like navigation tools — hold consequences for decision-making quality." The authors propose a few directions for study, and computer-aided representations are among them. In that regard, the sketches created by our 3D modelers (to make sense of raw descriptions and connotations) could provide insights for further examination.

VI. Related Literature

Here, we provide a more in-depth literature review of concepts relevant to this research.

Concreteness Effect and Emotion Words. The concreteness effect "refers to the observation that concrete nouns are processed faster and more accurately than abstract nouns in a variety of cognitive tasks" [92]. There are two well-known theories for explaining the effect's neuronal basis: the dual-coding theory, and the context availability theory. Jessen et al. [92] studies suggest a combination of both theories [92]. To account for experimental findings, both theories should link abstract words with experiential information [21]: Kousta et al. [21] study emotional content (a type of experiential information) to demonstrate that it plays a vital role in the processing and representation of abstract concepts.

Starting from the question "Are the concepts represented by emotion words different from abstract in memory?", Altarriba and Bauer [93] examine emotion concepts in three experiments. According to the authors, "although emotion words have often been included in the abstract stimuli in the literature, when rated on concreteness, imageability, and context availability they are different from abstract and concrete words". Altarriba and Bauer [93] results indicate that emotion words are more memorable and readily recalled than concrete and abstract words, and that concepts represented by emotion words are more imageable and are easier to find a context for than abstract words, although they are less concrete than abstract words. **Although we acknowledge these studies, we make a loose distinction between emotions and abstraction for simplicity since a deeper analysis falls out of the scope of our project.**

Text Comprehension. Research suggests that language comprehension involves sensorimotor representations; thus, Zwaan [20] reviews the literature on mental models focusing on how mental representations are constrained by linguistic and situational factors, which are then extended to include sensorimotor representations. Text Comprehension "is equivalent to the construction of multiple mental representations in working memory. (...) Mental representations include a text surface representation, a propositional representation, and a mental model. They are characterized by different forgetting rates. As speakers and authors omit information which can be easily completed by listeners and readers, text comprehension always includes inferences" [31]. With respect to mental models of the text content, "text comprehension can be characterized as the construction, evaluation, and (if needed) revision of a mental model of the subject matter described in the text" [31]. According to Schnotz [31], text meanings are constructed by the individual through an interaction between external information received through the text and internal information from the individual's prior knowledge" [31]

According to Butterfuss, Kim and Kendeou [94], reading involves three interrelated elements, all situated into a broader sociocultural context: 1) the reader, 2) text, and 3) the reading task. The authors provide considerations on individual differences in readering comprehension, and the importance of a readers' prior knowledge. They also consider the role of emotions in reading comprehension (information may elicit emotional responses). Within emotions, they point us to two key dimensions: valence and arousal. "Valence refers to whether the subjective experience of emotions is pleasant or unpleasant. Arousal refers to the level of physiological arousal and intent to engage in activity. These two dimensions of emotions may independently influence reading comprehension via attention, working memory, motivation, learning strategies, memory processes, and self-regulation" [94].

Networked Emotions and Mental Models. The term Networked Emotions (or "Messy Layers") takes into account the social nature of emotions and the messy layers of emotion and emotion regulation; it refers to the view of "emotions as multi-layered processes in which intraindividual processes are tightly coupled and often cannot be separated from interindividual processes" [22]. There are many instances where "regulation and elicitation can best be described by nested layers of feedback loops (...) Dealing with nested layers is messy because all layers can potentially influence emotional

components" [22]. Finally, according to Giaxoglou, Döveling, and Pitsillides [24], it "involves the mobilization of affect in online emotional cultures as a transmittable, spreadable, and self-contained resource, bringing out formerly privately shared emotions into online spaces and collective experience".

Nissenbaum and Shifman [95] present a cross-lingual study of memes to trace global and local expressive repertoires; and Flecha Ortiz and colleagues [96] investigate memes and collective coping theory, while discussing how memes can help to reinterpret a problematic situation. Continuing on coping theories, Schramm and Cohen [97] discuss emotion regulation and coping via media use.

Culture and Cognition. For Hutto et al. [98], sociocultural influences operate with respect to our explicitly formed and expressed beliefs and values but can additionally inform and infuse what we see and feel. Then, the authors [99] provide an interesting reflection on the production of the self and how continuous interaction with local cultural niches amplifies its scope through engagements with social media, ending up contributing to new ecologies of human existence. The authors [7] argue that we learn the shared habits and expectations of our culture through immersive participation in cultural practices that selectively shape attention and behaviour, a process by which the authors call "thinking through other minds" – finally, see [100] for more details in neuroscience research and culture.

Human Perception and 3D scenes. The computer modeling literature is active in producing insightful work on human perception (e.g., initiatives such as the Emotion Recognition Challenge [101]); as a review in computational sentiment analysis [75], and a survey on computational methods for modeling human perception of 3D scenes [102] show. The authors cover visual attention, 3D object quality perception, and material recognition. Forward, [103] review advances seeking to capture human efficiency in real-world scene and object perception, and [104] proposes a 3D modeling framework that uses visual attention characteristics to obtain compact models more adapted to human visual capabilities. Then, [105] offer insights into the development of applications in 3D knowledge of the scene, ranging from early stages of the 3D acquisition process to the higher-level tasks over 3D data. Finally, [106] provides a biologically constrained model of visual attention (with the capability of object recognition and localization) against large object variations of a visual search task in virtual reality.

Interestingly, [107] investigates the question of how to develop common sense in AI systems. Moving to semantic modeling, it could be used, for example, for large-scale scenes, automatically generating complex environments or supporting intelligent behavior on the virtual scenes, semantic rendering, and adaptive visualization of complex 3D objects [108]. Switching gears to narratives, Ong et al. [109] review time-series emotion recognition and time-series approaches in affective computing; finally, they introduce the Stanford Emotional Narratives Dataset (SENDv1), a set of rich, multimodal videos of self-paced, unscripted, emotional narratives, annotated for emotional valence over time. Finally, see [110] for a context-aware emotion recognition framework that combines four contexts: multimodal emotion recognition based on facial expression, facial landmarks, gesture, and gait.

Skurka and Nabi [111] discuss four traditions of emotion theory and highlight how digital spaces can contribute to emotional arousal and impact. Then, [112] focuses on the cognitive science of human variation in the field of spatial navigation since studies either using the real world or virtual reality show that there are significant individual differences in navigation competencies. Aiming to help researchers and designers develop emotionally interactive devices or designs, [113] examine emotional interactions between humans and deformable objects; they investigate how the design of a flexible display (depicted in 3D images in which an object is bent at different axes) interacts with emotion. Thinking of spatial skills and objects in 3D, Munns et al. [114] present an approach for developing computer-based tests of spatial skills and illustrate it by creating a test of the ability to visualize cross-Sections of 3D objects.

Nissenbaum and Shifman [95] present a cross-lingual study of memes to trace global and local expressive repertoires; and Flecha Ortiz and colleagues [96] investigate memes and collective coping theory, while discussing how memes can help to reinterpret a problematic situation. Continuing on coping theories, Schramm and Cohen [97] discuss emotion regulation and coping via media use. Finally, we conclude with considerations on empathy: Zaki et al. [115] reflect on a lack of a consistent demonstration of a correspondence between affective empathy (perceivers' experience of social targets' emotions) and empathic accuracy (perceivers' ability to accurately assess targets' emotions) – important since theories suggest that affective empathy should contribute to empathic accuracy. Their findings suggest that perceivers' self-reported affective empathy can predict their empathic accuracy, but only when targets' expressivity allows their thoughts and feelings to be read.

VII. Conclusion

In Active Threat Response Training, you are likely urged to comprehend the meaning behind your perception. What if you wanted to somehow architect these skills into a machine? Seeking to investigate what a holistic encoding of abstract and emotion-rich contexts could look like for an emotion-driven AI system, we created a multi-phase project to examine how abstraction and emotions travel different spaces.

We detailed our project's phases 1 and 2: the Phase 1 team provides raw textual image descriptions to the Phase 2 team, which is responsible for turning a description into a 3D scene. We presented our methods for creating textual descriptions from memes and a dataset that focuses on the image's observer. We also show a sample of 3D scenes' images along with corresponding raw descriptions and the Gateways diagram, designed to help us understand the 3D modelers' decision-making. We identified applications for social impact but will expand on that in future work. We hope our dataset, raw descriptions, and Gateways diagram can provide insights

into exploring the concreteness effect in connection with sensorimotor representations.

Alt-text and long descriptions are essential for conveying the visual aspects of images to individuals with print disabilities, as we have previously discussed. Memes and other humorous images are key components of digital culture, fostering connections among people. If a framework can be provided that methodically and thoughtfully takes into account the *assumed elements* and guides the creation of image descriptions in reference to images with emotional/hidden meanings, such as memes, it will better include those who cannot see the visuals. Our collected images focus on humorous/entertaining aspects, but we hope our methods can inform approaches to expand on other themes.

Given emotions' investigation challenges, we are identifying processes to ensure objectivity and map how a message travels through spaces. As we do so, challenging questions emerge, and we hypothesize they will bring insights into research in emotions and how to build emotion-aware AI systems and assistive technologies. For example, how informative would it be if an *emotion-driven* AI system outputted its decision log on missing and *assumed elements* in a narrative-like sequence of pictures (or 3D scenes) and text?

To conclude, the work [116] describes a computational model that uses multiple representations in problem-solving. The model's behavior is illustrated by simulating the "cognitive and perceptual processes of an economics expert as he teaches some well-learned economics principles while drawing a graph on a black-board". It would be interesting to combine [116] with our methods and the Gateways Diagram to simulate a 3D modeler creating 3D scenes from raw descriptions.

ACKNOWLEDGMENTS

This work would not have been possible without the support from Grinnell College's Harris Faculty Fellowship and the Mentored Advanced Project (MAP) program.

Finally, we would like to thank our reviewers for their thoughtful comments.

REFERENCES

[1] J. Chen, E. Berman, M. Noda, K. Shermak, Z. Ye, D. Rothfusz, and F. Eliott, "How do abstraction and emotions travel different spaces?" in *Proc. of The Tenth International Conference on Human and Social Analytics HUSO*. IARIA, 2024.

[2] R. Dawkins, "The selfish gene," in *The selfish gene*, 1976, pp. 224–p.

[3] B. Bettin, A. Sarabia, M. C. Gonzalez, I. Gatti, C. Magnan, N. Murav, R. Vanden Heuvel, D. McBride, and S. Abraham, "Say what you meme: Exploring memetic comprehension among students and potential value of memes for cs education contexts," in *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 1*, 2023, pp. 416–429.

[4] J. Scott, *A Dictionary of Sociology (Connotative versus denotative meaning entry)*. Oxford University Press, 2015.

[5] Merriam-Webster, "Connotation vs. denotation: Literally, what do you mean?" 2024. [Online]. Available: https://www.merriam-webster.com/grammar/connotation-vs-denotation-literally-what-do-you-mean

[6] A. M. Seidenfeld, S. R. Johnson, E. W. Cavadel, and C. E. Izard, "Theory of mind predicts emotion knowledge development in head start children," *Early Education and Development*, vol. 25, no. 7, pp. 933–948, 2014.

[7] S. P. L. Veissière, A. Constant, M. J. D. Ramstead, K. J. Friston, and L. J. Kirmayer, "Thinking through other minds: A variational approach to cognition and culture," *Behavioral and Brain Sciences*, vol. 43, p. e90, 2020.

[8] K. L. van den Broek, J. Luomba, J. van den Broek, and H. Fischer, "Evaluating the application of the mental model mapping tool (m-tool)," *Frontiers in Psychology*, vol. 12, p. 761882, 2021.

[9] B. Stangl, "Emotional mental models," in *Encyclopedia of the Sciences of Learning*. Springer, 2012.

[10] D. S. Robert W. Andrews, J. Mason Lilly and K. M. Feigh, "The role of shared mental models in human-ai teams: a theoretical review," *Theoretical Issues in Ergonomics Science*, vol. 24, no. 2, pp. 129–175, 2023.

[11] S. Lodha and R. Gupta, "Irrelevant angry, but not happy, faces facilitate response inhibition in mindfulness meditators," *Current Psychology*, vol. 43, no. 1, pp. 811–826, 2024.

[12] S. Gadanho, "Reinforcement learning in autonomous robots: an empirical investigation of the role of emotions," Ph.D. dissertation, U. of Edinburgh. College of Science and Engineering. School of Informatics., 1999.

[13] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction (2nd ed.)*. MIT press, (1992), 2018.

[14] X. Yu, R. Morri, and F. Eliott, "Eda, an empathy-driven computational architecture," in *Proceedings of the Ninth Goal Reasoning Workshop at ACS.*, 2021.

[15] F. Eliott and C. Ribeiro, "Moral behavior and empathy modeling through the premise of reciprocity," in *Proc. of The First International Conference on Human and Social Analytics HUSO*. IARIA, 2015.

[16] ——, "Emergence of cooperation through simulation of moral behavior," in *Hybrid Artificial Intelligent Systems. HAIS 2015: 10th I. Conf. on Hybrid Artificial Intelligence Systems, Bilbao, Spain. Lecture Notes in Artificial Intelligence*, vol. 9121. Springer International Pub., 2015, pp. 200–212.

[17] M. G. Mattar, J. E. Fan, W. K. Vong, and L. Wong, "How does the mind discover useful abstractions?" in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 45, no. 45, 2023.

[18] S. K. Reed, "A taxonomic analysis of abstraction," *Perspectives on Psychological Science*, vol. 11, no. 6, pp. 817–837, 2016. [Online]. Available: http://www.jstor.org/stable/26358684

[19] M. K. Ho, D. Abel, T. L. Griffiths, and M. L. Littman, "The value of abstraction," *Current opinion in behavioral sciences*, vol. 29, pp. 111–116, 2019.

[20] R. A. Zwaan, "Situation models, mental simulations, and abstract concepts in discourse comprehension," *Psychonomic bulletin & review*, vol. 23, pp. 1028–1034, 2016.

[21] S.-T. Kousta, G. Vigliocco, D. P. Vinson, M. Andrews, and E. Del Campo, "The representation of abstract words: why emotion matters." *Journal of Experimental Psychology: General*, vol. 140, no. 1, p. 14, 2011.

[22] A. Kappas, "Social regulation of emotion: messy layers," *Frontiers in psychology*, vol. 4, p. 51, 2013.

[23] T. Jiang, H. Li, and Y. Hou, "Cultural differences in humor perception, usage, and implications," *Frontiers in psychology*, vol. 10, p. 438919, 2019.

[24] K. Giaxoglou, K. Döveling, and S. Pitsillides, "Networked emotions: Interdisciplinary perspectives on sharing loss online," pp. 1–10, 2017.

[25] A. Halversen and B. E. Weeks, "Memeing politics: Understanding political meme creators, audiences, and consequences on social media," *Social Media + Society*, vol. 9, no. 4, p. 20563051231205588, 2023.

[26] Round Table team, "Round table on information access for people with print disabilities," 2024. [Online]. Available: https://printdisability.org/

[27] C. F. Karbowski, "See3d: 3d printing for people who are blind." *Journal of Science Education for Students with Disabilities*, vol. 23, no. 1, p. n1, 2020.

[28] M. University, "Accessible graphics hub," 2024. [Online]. Available: https://accessiblegraphics.org/

[29] E. Swaim and F. Eliott, "Complex behavior vs. design-interpreting ai: Reminders from synthetic psychology," in *Proc. of The 9th International Conference on Human and Social Analytics HUSO*. IARIA, 2023.

[30] Posit, "R shiny application." [Online]. Available: https://shiny.posit.co/

[31] W. Schnotz, *Comprehension of Text*. Cambridge University Press, 2023, p. 63–86.

[32] Accessible Publishing Contributers, ""accessible publishing. an online portal featuring information and resources for the advancement and development of accessible publishing in canada and beyond."." [Online]. Available: https://www.accessiblepublishing.ca/a-guide-to-image-description/#terms

[33] J. T. Nganji, M. Brayshaw, and B. Tompsett, "Describing and assessing image descriptions for visually impaired web users with idat," in *Proceedings of the Third International Conference on Intelligent Human Computer Interaction (IHCI 2011), Prague, Czech Republic, August, 2011*. Springer, 2012, pp. 27–37.

[34] V. Lewis, ""veronica with four eyes"," 2024. [Online]. Available: https://veroniiiica.com/how-to-write-alt-text-for-memes/

[35] J. J. Gibson, "The senses considered as perceptual systems." 1966.

[36] ——, *The ecological approach to visual perception: classic edition*. Psychology press, [1979], 2014.

[37] K. Zmanovskaia, "Cats photoshopped into food are so cute you could just eat them up, by emma taggart." [Online]. Available: https://mymodernmet.com/cats-in-food-photoshop-funny/

[38] J. Zheng, M. Zhou, J. Mo, and A. Tharumarajah, "Background and foreground knowledge in knowledge management," in *International Working Conference on the Design of Information Infrastructure Systems for Manufacturing*. Springer, 2000, pp. 332–339.

[39] F. Nickols, "The tacit and explicit nature of knowledge: The knowledge in knowledge management," in *The knowledge management yearbook 2000-2001*. Routledge, 2013, pp. 12–21.

[40] M. Llorens-Gámez, J. L. Higuera-Trujillo, C. S. Omarrementeria, and C. Llinares, "The impact of the design of learning spaces on attention and memory from a neuroarchitectural approach: A systematic review," *Frontiers of Architectural Research*, vol. 11, no. 3, pp. 542–560, 2022.

[41] J. Leshin, M. J. Carter, C. M. Doyle, and K. A. Lindquist, "Language access differentially alters functional connectivity during emotion perception across cultures," *Frontiers in Psychology*, vol. 14, p. 1084059, 2024.

[42] B. Mesquita, *Between us: How cultures create emotions*. WW Norton & Company, 2022.

[43] Z. H. Pugh, S. Choo, J. C. Leshin, K. A. Lindquist, and C. S. Nam, "Emotion depends on context, culture and their interaction: evidence from effective connectivity," *Social Cognitive and Affective Neuroscience*, vol. 17, no. 2, pp. 206–217, 07 2021. [Online]. Available: https://doi.org/10.1093/scan/nsab092

[44] L. Soni, A. Kaur, and A. Sharma, "A review on different versions and interfaces of blender software," in *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 2023, pp. 882–887.

[45] T. M. Takala, M. Mäkäräinen, and P. Hämäläinen, "Immersive 3d modeling with blender and off-the-shelf hardware," in *2013 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, 2013, pp. 191–192.

[46] B. Tversky, "Visualizing thought," in *Handbook of human centric visualization*. Springer, 2014, pp. 3–40.

[47] J. R. Zadra and G. L. Clore, "Emotion and perception: The role of affective information," *Wiley interdisciplinary reviews: cognitive science*, vol. 2, no. 6, pp. 676–685, 2011.

[48] E. L. Cohen and J. G. Myrick, "Emotions and technological affordances," *Emotions in the Digital World: Exploring Affective Experience and Expression in Online Interactions*, p. 32, 2023.

[49] S. Kriz and M. Hegarty, "Top-down and bottom-up influences on learning from animations," *International Journal of Human-Computer Studies*, vol. 65, no. 11, pp. 911–930, 2007.

[50] M. Hegarty, "Diagrams in the mind and in the world: Relations between internal and external visualizations," in *Diagrammatic Representation and Inference: Third International Conference, Diagrams 2004, Cambridge, UK, March 22-24, 2004. Proceedings 3*. Springer, 2004, pp. 1–13.

[51] ——, "The cognitive science of visual-spatial displays: Implications for design," *Topics in cognitive science*, vol. 3, no. 3, pp. 446–474, 2011.

[52] M. Hegarty and M.-A. Just, "Constructing mental models of machines from text and diagrams," *Journal of memory and language*, vol. 32, no. 6, pp. 717–742, 1993.

[53] L. M. Padilla, S. H. Creem-Regehr, M. Hegarty, and J. K. Stefanucci, "Decision making with visualizations: a cognitive framework across disciplines," *Cognitive research: principles and implications*, vol. 3, no. 1, pp. 1–25, 2018.

[54] S. Pinker, "A theory of graph comprehension," in *Artificial intelligence and the future of testing*. Psychology Press, 2014, pp. 73–126.

[55] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *Proceedings of international conference on intelligence analysis*, vol. 5. McLean, VA, USA, 2005, pp. 2–4.

[56] J. Pearson, "The human imagination: the cognitive neuroscience of visual mental imagery," *Nature reviews neuroscience*, vol. 20, no. 10, pp. 624–634, 2019.

[57] "Machine learning meme." [Online]. Available: https://www.reddit.com/r/ProgrammerHumor/comments/jxyawv/machine_learning/

[58] M. Zins, "The future we all face." [Online]. Available: https://www.cartoonmovement.com/cartoon/future-we-all-face

[59] "Butterfly meme, by i love killing flies." [Online]. Available: https://www.facebook.com/ILoveKillingFlies/

[60] "Covid memes: Why we're using laughter to get us through a pandemic, by karine bengualid." [Online]. Available: https://copyhackers.com/2020/06/covid-memes/

[61] "Awkward half-cat loafing on the stairs sparks photoshop battle no one expected, by andrea romano." [Online]. Available: https://mashable.com/article/awkward-half-cat-photoshop-battle

[62] "These cute illustrations prove cats are just funny little shapes." [Online]. Available: https://www.buzzfeed.com/pablovaldivia/silly-cat-drawings

[63] G. Giorgi, "Methodological directions for the study of memes," in *Handbook of research on advanced research methodologies for a digital society*. IGI Global, 2022, pp. 627–663.

[64] L. Cochrane, A. Johnson, A. Lay, and G. Helmandollar, ""one does not simply categorize a meme": A dual classification system for visual-textual internet memes," *Proceedings of the Linguistic Society of America*, vol. 7, no. 1, pp. 5260–5260, 2022.

[65] L. Marti, S. Wu, S. T. Piantadosi, and C. Kidd, "Latent Diversity in Human Concepts," *Open Mind*, vol. 7, pp. 79–92, 03 2023.

[66] I. Lozano-Palacio and F. J. R. de Mendoza Ibáñez, *Modeling Irony : A Cognitive-Pragmatic Account*. John Benjamins, 2022.

[67] Genie team, "Genie: Generative interactive environments," 2024. [Online]. Available: https://sites.google.com/view/genie-2024/

[68] OpenAI team, "Chatgpt can now see, hear and speak," 2024. [Online]. Available: https://openai.com/blog/chatgpt-can-now-see-hear-and-speak

[69] C. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M. Liu, and T. Lin, "Magic3d: High-resolution text-to-3d content creation," in *IEEE Conf. on Comp. Vision and Pattern Recog. (CVPR)*, 2023.

[70] G. Chechik, R. Gal, and Y. Atzmon, "Generative ai research spotlight: Personalizing text-to-image models," 2024. [Online]. Available: https://developer.nvidia.com/blog/generative-ai-research-spotlight-personalizing-text-to-image-models/

[71] Adobe team, "Adobe firefly," 2024. [Online]. Available: https://www.adobe.com/products/firefly.html

[72] OpenAI team, "Creating video from text," 2024. [Online]. Available: https://openai.com/sora

[73] "Genz 4 meme," https://chatgpt.com/g/g-OCOyXYJjW-genz-4-meme, 2024, hosted on ChatGPT by OpenAI.

[74] M. Mitchell, A. B. Palmarini, and A. Moskvichev, "Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks," *arXiv preprint arXiv:2311.09247*, 2023.

[75] Y. Ophir and D. Walter, "Computational sentiment analysis," *Emotions in the Digital World: Exploring Affective Experience and Expression in Online Interactions*, p. 114, 2023.

[76] A. Aggarwal, V. Sharma, A. Trivedi, M. Yadav, C. Agrawal, D. Singh, V. Mishra, and H. Gritli, "Two-way feature extraction using sequential and multimodal approach for hateful meme classification," *Complexity*, vol. 2021, pp. 1–7, 2021.

[77] P. Behera, A. Ekbal *et al.*, "Only text? only image? or both? predicting sentiment of internet memes," in *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, 2020, pp. 444–452.

[78] M. Buxbaum, H. F. Pedersen Ed D *et al.*, "What do you meme? meme humor comprehension in adolescents with language disorder or hearing loss," *The Journal of Special Education Apprenticeship*, vol. 11, no. 1, p. 6, 2022.

[79] U. Akram, J. Drabble, G. Cau, F. Hershaw, A. Rajenthran, M. Lowe, C. Trommelen, and J. G. Ellis, "Exploratory study on the role of

emotion regulation in perceived valence, humour, and beneficial use of depressive internet memes in depression," *Scientific reports*, vol. 10, no. 1, p. 899, 2020.

[80] A. Zeman, "Aphantasia and hyperphantasia: exploring imagery vividness extremes," *Trends in Cognitive Sciences*, 2024.

[81] S. O. Mathiesen and A. Canossa, "If you can't beat them, join them: How text-to-image tools can be leveraged in the 3d modelling process," in *HCI International 2023 – Late Breaking Papers*. Cham: Springer Nature Switzerland, 2023, pp. 162–181.

[82] S. Reinders, "Accessible interactive 3d models for blind and low-vision people," *ACM SIGACCESS Accessibility and Computing*, no. 129, pp. 1–7, 2021.

[83] C. C. Spector, "Remediating humor comprehension deficits in language-impaired students," *Language, speech, and hearing services in schools*, vol. 23, no. 1, pp. 20–27, 1992.

[84] T. Grandin, "Emergence: Labeled autistic (with margaret scariano)," *Arena Press*, 1986.

[85] ——, *Thinking in pictures: My life with autism*. Vintage, 1995.

[86] C. C. Spector, *As far as words go: activities for understanding ambiguous language and humor*. Brookes Publishing, 2009.

[87] ——, "Just for laughs: Understanding multiple meanings in jokes," 1995.

[88] ——, "Saying one thing, meaning another: Activities for clarifying ambiguous language," 1997.

[89] ——, "Between the lines enhancing inferencing skills," 2006.

[90] Y. Bai and W. Bainbridge, "Diagnostic images for alzheimer's disease show distinctions in biomarker status and scene-related functional activity between patients and healthy controls," *Journal of Vision*, vol. 23, no. 9, pp. 5600–5600, 2023.

[91] F. A. Csaszar, N. Hinrichs, and M. Heshmati, "External representations in strategic decision-making: Understanding strategy's reliance on visuals," *Strategic Management Journal*, vol. n/a, no. n/a, 2024.

[92] F. Jessen, R. Heun, M. Erb, D.-O. Granath, U. Klose, A. Papassotiropoulos, and W. Grodd, "The concreteness effect: Evidence for dual coding and context availability," *Brain and language*, vol. 74, no. 1, pp. 103–112, 2000.

[93] J. Altarriba and L. M. Bauer, "The distinctiveness of emotion concepts: A comparison between emotion, abstract, and concrete words," *The American journal of psychology*, pp. 389–410, 2004.

[94] R. Butterfuss, J. Kim, and P. Kendeou, "Reading comprehension." *Grantee Submission*, 2020.

[95] A. Nissenbaum and L. Shifmancohen2023emotions, "Meme templates as expressive repertoires in a globalizing world: A cross-linguistic study," *Journal of Computer-Mediated Communication*, vol. 23, no. 5, pp. 294–310, 2018.

[96] J. A. Flecha Ortiz, M. A. Santos Corrada, E. Lopez, and V. Dones, "Analysis of the use of memes as an exponent of collective coping during covid-19 in puerto rico," *Media International Australia*, vol. 178, no. 1, pp. 168–181, 2021.

[97] H. Schramm and E. L. Cohen, "Emotion regulation and coping via media use," *The international encyclopedia of media effects*, pp. 1–9, 2017.

[98] D. D. Hutto, S. Gallagher, J. Ilundáin-Agurruza, and I. Hipólito, *Culture in Mind – An Enactivist Account: Not Cognitive Penetration but Cultural Permeation*, ser. Current Perspectives in Social and Behavioral Sciences. Cambridge University Press, 2020, p. 163–187.

[99] *The Situated Brain: Introduction*, ser. Current Perspectives in Social and Behavioral Sciences. Cambridge University Press, 2020, p. 159–272.

[100] S. Han and G. Northoff, *Cultural Priming Effects and the Human Brain*, ser. Current Perspectives in Social and Behavioral Sciences. Cambridge University Press, 2020, p. 223–243.

[101] Odyssey, "Emotion recognition challenge," 2024. [Online]. Available: https://www.odyssey2024.org/emotion-recognition-challenge

[102] Z. C. Yildiz, A. Bulbul, and T. Capin, "Modeling human perception of 3d scenes," in *Intelligent Scene Modeling and Human-Computer Interaction*. Springer, 2021, pp. 67–88.

[103] T. Lauer and M. L.-H. Võ, "The ingredients of scenes that affect object search and perception," *Human perception of visual information: Psychological and computational perspectives*, pp. 1–32, 2022.

[104] M. Chagnon-Forget, G. Rouhafzay, A.-M. Cretu, and S. Bouchard, "Enhanced visual-attention model for perceptually improved 3d object modeling in virtual environments," *3D Research*, vol. 7, pp. 1–18, 2016.

[105] M. Poggi and T. B. Moeslund, "Computer vision for 3d perception and applications," p. 3944, 2021.

[106] A. Jamalian, F. Beuth, and F. H. Hamker, "The performance of a biologically plausible model of visual attention to localize objects in a virtual reality," in *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*. Springer, 2016, pp. 447–454.

[107] Y. Zhu, T. Gao, L. Fan, S. Huang, M. Edmonds, H. Liu, F. Gao, C. Zhang, S. Qi, Y. N. Wu *et al.*, "Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense," *Engineering*, vol. 6, no. 3, pp. 310–345, 2020.

[108] J. Divya Udayan and H. Kim, "Semantic modeling and rendering," in *Intelligent Scene Modeling and Human-Computer Interaction*. Springer, 2021, pp. 105–127.

[109] D. C. Ong, Z. Wu, Z.-X. Tan, M. Reddan, I. Kahhale, A. Mattek, and J. Zaki, "Modeling emotion in complex stories: the stanford emotional narratives dataset," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 579–594, 2019.

[110] D. Yang, S. Huang, S. Wang, Y. Liu, P. Zhai, L. Su, M. Li, and L. Zhang, "Emotion recognition for multiple context awareness," in *European Conference on Computer Vision*. Springer, 2022, pp. 144–162.

[111] C. Skurka and R. L. Nabi, "Perspectives on emotion in the digital age." 2023.

[112] N. S. Newcombe, M. Hegarty, and D. Uttal, "Building a cognitive science of human variation: Individual differences in spatial navigation," pp. 6–14, 2023.

[113] J. M. Lee, J. Baek, and D. Y. Ju, "Anthropomorphic design: emotional perception for deformable object," *Frontiers in psychology*, vol. 9, p. 1829, 2018.

[114] M. E. Munns, C. He, A. Topete, and M. Hegarty, "Visualizing cross-sections of 3d objects: Developing efficient measures using item response theory," *Journal of Intelligence*, vol. 11, no. 11, 2023. [Online]. Available: https://www.mdpi.com/2079-3200/11/11/205

[115] J. Zaki, N. Bolger, and K. Ochsner, "It takes two: The interpersonal nature of empathic accuracy," *Psychological science*, vol. 19, no. 4, pp. 399–404, 2008.

[116] H. J. Tabachneck-Schijf, A. M. Leonardo, and H. A. Simon, "Camera: A computational model of multiple representations," *Cognitive science*, vol. 21, no. 3, pp. 305–350, 1997.

[117] Merriam-Webster, "Accuracy," 2023. [Online]. Available: https://www.merriam-webster.com/dictionary/accuracy

[118] ——, "Reliable," 2024. [Online]. Available: https://www.merriam-webster.com/dictionary/reliable

[119] ——, "Consistent," 2024. [Online]. Available: https://www.merriam-webster.com/dictionary/consistent

[120] ——, "Object," 2023. [Online]. Available: https://www.merriam-webster.com/dictionary/object

[121] ——, "Faithful," 2023. [Online]. Available: https://www.merriam-webster.com/dictionary/faithfulness

[122] ——, "Abstract," 2023. [Online]. Available: https://www.merriam-webster.com/dictionary/abstract

[123] A. Damasio and G. B. Carvalho, "The nature of feelings: evolutionary and neurobiological origins," *Nature reviews neuroscience*, vol. 14, no. 2, pp. 143–152, 2013.

[124] X. Hu and M. Twidale, "A scoping review of mental model research in hci from 2010 to 2021," in *HCI International 2023 – Late Breaking Papers*. Cham: Springer Nature Switzerland, 2023, pp. 101–125.

[125] B. Tversky, "Cognitive maps, cognitive collages, and spatial mental models," in *European conference on spatial information theory*. Springer, 1993, pp. 14–24.

[126] M. Hegarty, "Mechanical reasoning by mental simulation," *Trends in cognitive sciences*, vol. 8, no. 6, pp. 280–285, 2004.

[127] M. Hegarty and D. Waller, "A dissociation between mental rotation and perspective-taking spatial abilities," *Intelligence*, vol. 32, no. 2, pp. 175–191, 2004.

[128] M. Hegarty, "Chapter 7 - components of spatial intelligence," in *The Psychology of Learning and Motivation*, ser. Psychology of Learning and Motivation. Academic Press, 2010, vol. 52, pp. 265–297.

APPENDIX A: GLOSSARY

We created a glossary to ensure consistency in our communication and processes. Overall, terms are sorted for better understanding instead of alphabetically.

We identified at least three concepts for communicating connotative, emotion-rich messages in digital spaces, which are the extent to which a message's material representation is:

1) **Accurate/Accuracy.** A "conformity to truth or to a standard or model" [117]. In our context, if something captures the source's explicit and concrete elements.
2) **Reliable.** "suitable or fit to be relied on." [118] Here, if a detailed textual description reliably captures the images' abstraction and contextual linkages needed to retrieve its meaning (see "Faithfulness" below).
3) **Consistent.** "marked by harmony, regularity, or steady continuity: free from variation or contradiction" [119]. Here, if a message remains *consistent* with the original source, in spite of traveling through various spaces.

Terms More Related to the task of building the 3D Scenes:

- **3D Models.** Refer to the 3D modeling process. Once the model passes the two gateways (figure 6), it is complete/finished, and we call it a **3D Scene**. A modeler's goal is not to make fancy 3D scenes; they stop modeling once they determine a model matches their mental models triggered by the raw description.
- **Concrete Elements.** Elements that add specific and objective visual elements to an object, e.g., modeling a cat sitting on a chair.
- **Object.** "Something material that may be perceived by the senses" [120], e.g., a cat or a person.
- **Modeled Emotion**: Emotion that the modeler seeks to model onto the concrete objects in the scene. E.g., adding expressive features to facial expressions so that an emotion can be visually seen on the object.
- **Subject.** Concrete, material element(s) of all elements in the scene that the modeler identifies as a dominant, primary component of the entire scene.
- **Character.** Object that, from the description, seems to express or elicit emotions.
- **Observer.** The perspective from a person viewing from outside the image or 3D scene.
- **Participant Observer.** If the modeler identifies, from the raw description, that the scene must allow an observer to merge with the scene so that the observer is also a participant (e.g., "Hand with Reflecting Sphere" by Maurits C. Escher). It corresponds to our dataset attribute *Outward (ad hoc) participant/observer*.
- **Intended Observer's Emotional Response.** The emotions modelers intend to elicit in the observer by looking at the 3D scene. In humor, many times, the intended emotion conflicts with modeled emotions (e.g., a scene of a cat not enjoying a bath is likely to be funny to an observer whose perspective is from the outside of the 3D scene).
- **Faithfulness**: The extent to which the 3D model is 'true' to the modeler's emotional mental models of the scene triggered by the raw description. We use the definition: "true to the facts, to a standard, or to an original" [121], which are the modeler's emotional mental models. Once it is 'true', the 3D model passes the *faithfulness* gateway, as shown in our Gateways diagram. Here, "reliable" relies more on the concrete source (e.g., image), whereas "faithfulness" on the co-creation of someone (modeler) blending together pieces from a concrete source (descriptions) and generated mental models.
- **3D Model Accuracy.** The extent to which the 3D model reflects the explicit and concrete elements triggered by the raw description in the modeler's mental models. Once it reflects those elements, the 3D model passes the *accuracy* gateway, as shown in our Gateways diagram.
- **Missing Elements.** Elements a modeler identifies to be missing from the raw description. Then, modelers use their knowledge and experiences to make assumptions, turning missing elements into assumed elements to pass through both gateways: *accuracy* and *faithfulness*.
- **Assumed Elements.** Elements intentionally added by modelers to the Raw 3D Model to "fill in the gap" left by the missing elements. That enables modelers to shape the 3D model to match their mental models of the scene triggered by the raw description.

Other relevant terms:

- **Abstract.** "Expressing a quality apart from an object" [122]. We use abstraction as an umbrella term that intercepts connotative meanings and emotion and mental models.
- **Connotative vs. Denotative Meaning.** "Connotative meaning refers to the associations, overtones, and feel that a concept has, rather than what it refers to explicitly (or denotes, hence denotative meaning). Two words with the same reference or definition may have different connotations" [4].
- **Explicit, Implicit, and Tacit Knowledge.** "When knowledge has been articulated, then it is explicit knowledge. Otherwise, another question is raised: Can it be articulated? If the answer is yes, then it is implicit knowledge. If the answer is no, then it is tacit knowledge" [38].
- **Emotion-rich message.** Anything that conveys emotional messages that human senses can perceive.
- **Emotions and Feelings.** Both are key concepts for homeostatic regulation: "Feelings are mental experiences of body states. They signify physiological need (for example, hunger), tissue injury (for example, pain), optimal function (for example, well-being), threats to the organism (for example, fear or anger) or specific social interactions (for example, compassion, gratitude or love)". Whereas "Emotions include disgust, fear, anger, sadness, joy, shame, contempt, pride, compassion and admiration, and they are mostly triggered by the perception or recall of exteroceptive stimuli" [123]. Emotions "regulate

social interaction and in extension, the social sphere. In turn, processes in the social sphere regulate emotions of individuals and groups" [22].

- **Emotional Mental Models.** Cover emotions and feelings connected to mental models. Hu and Twidale [124] provide a broad definition of mental models: they "refer to humans' internal representations of the external world that derive from their perception, memories, knowledge, and causal beliefs". As Hu and Twidale [124], we acknowledge that the term "mental models" has a multidimensional nature, and below we provide more context for Emotional Mental Models: "Mental models cause certain expectations/thoughts of how things should look like/work and connect certain emotions with this. Consequently, a mental model is a cognitive and an emotional framework in the brain, influenced by person's personality (genes) and the environment including social variables" [9].
- **Mental Models.** "internal representations of the external world consisting of causal beliefs that help individuals deduce what will happen in a particular situation" [8]. For simplicity, we use 'mental models' as an umbrella term that covers terms such as spatial mental models and mental representations of environments or 'cognitive collages' [125]. Although that simplification is not ideal and considerations on mental simulation and mechanical reasoning [126] are extremely relevant, such an examination falls out of the scope of this manuscript. Likewise, considerations on a distinction between "mental abilities that require a spatial transformation of a perceived object (e.g., mental rotation) and those that involve imagining how a scene looks like from different viewpoints (e.g., perspective taking)" [127].
- **Networked Emotions ("Messy Layers").** Takes into account the social nature of emotions and the messy layers of emotion and emotion regulation. It refers to the view of "emotions as multi-layered processes in which intraindividual processes are tightly coupled and often cannot be separated from interindividual processes" [22]. It "involves the mobilization of affect in online emotional cultures as a transmittable, spreadable, and self-contained resource, bringing out formerly privately shared emotions into online spaces and collective experience" [24].
- **Humor.** "results when the incongruous is resolved (i.e., the punchline is seen to make sense at some level with the earlier information in the joke). Lacking a resolution the individual does not "get" the joke, is puzzled or even frustrated. The resolution phase is a form of problem solving, an attempt to draw information or inferences that make a link between the initial body of the joke or cartoon and its ending" [83].
- **Irony.** Is "determined by the attitudinal element arising from the clash between an epistemic and an observable scenario". We consider verbal and situational irony as different materializations of the same phenomenon: "In both cases, the epistemic scenario is drawn from the speaker's certainty about a state of affairs (be it formed through an echo or not), and the observable scenario from the situation that is evident to the speaker" [66].
- **Memes.** "A form of media communicating a thought or idea through some shared understanding" [3].
- **Image Description.** It is an umbrella term for image descriptions in a textual form.
- **Detailed Description.** Our detailed descriptions aim to fully describe images that have complex, abstract messages.
- **Alt-text.** "Alt-text, also known as alternative text, offers textual description of images. These text descriptions are visually hidden but when a blind or visually impaired reader encounters an image while using their screen reader, the alt-text will be read out. Descriptions are generally concise" [32]. They are "text-based descriptions of visual details in an image written primarily for people who are visually impaired (inclusive of blind/low vision)" [34].
- **Caption.** "A caption is a visible text component which accompanies an image and provides additional information. It may describe the image briefly and/or give contextual information about the source. It does not usually describe the image in great detail but instead, works in conjunction with the image" [32].
- **Sense-Making Tasks.** They "consist of information gathering, re-representation of the information in a schema that aids analysis, the development of insight through the manipulation of this representation, and the creation of some knowledge product or direct action based on the insight. In a formula Information $\rightarrow$ Schema $\rightarrow$ Insight $\rightarrow$ Product" [55], and the re-representation may be in the modeler's mind, written or drawn, or even digitally represented.
- **Spatial thinking.** It "involves thinking about the shapes and arrangements of objects in space and about spatial processes, such as the deformation of objects, and the movement of objects and other entities through space. It can also involve thinking with spatial representations of nonspatial entities". And spatial intelligence "can be defined as adaptive spatial thinking" [128].

APPENDIX B: THE OBSERVER-CENTERED DATASET ATTRIBUTES

The Concrete Design dimension focuses on concrete characteristics, and it splits into five categories and 14 attributes. The Blend dimension has six categories (two shared with the Emotional design) and 11 attributes (4 from the shared categories). Due to better alignment, we depict the shared categories within the Emotional Design dimension, which focuses on networked emotions and has three categories and 5 attributes (4 shared). In Appendix B, we detail the attributes.

The Concrete Design Dimension categories and attributes are as follows:

1) Logistics: attributes related to handling an image.
   (a) Image ID: image's numbered identification; format: image_#number.
   (b) Source. The memes' source, if available (N/A otherwise).
   (c) Date Processed. The most recent date (month/day/year) an image's attributes were updated/completed.
   (d) Tags: three words/short sentences that help to identify an image.
2) Concrete Elements: image's main subjects.
   (a) Subject(s): those are the concrete, material element(s) of all elements in the scene that have been identified as a dominant(s), primary component(s) of the entire image. There are no fixed attribute options, as they are meant to provide context about what the image is about. E.g., "cat", "person", "cactus that looks like a cat".
   (b) Subjects' number: registers how many subjects are the focus of the image; select a number between $1-10$, "M" if there are more than ten subjects in the image focus, and "N/A" either if there is no clear focus or if subjects are absent.
3) Distortion: if the image shows any distortion.
   (a) Synthetic Component: whether an image has been clearly altered to achieve a certain effect (such as adding a drawing on top of a picture). Select one of the entries: Absent (it does not seem to have been modified in any way), Edited (has clearly been altered), Live (it looks like being modified in real-time while it is being created, similar to M. C. Escher's *Drawing Hands*).
   (b) Reality Divergence or distortion: whether an image deviates from the expected reality in which it is presented. This includes instances where there are synthetic components or staged appearances of objects or creatures performing actions that are not possible in reality. This attribute is binary, with "True" indicating a divergence from reality and "False" indicating that the image adheres to reality. While non-photographic images such as drawings or cartoons may have more flexibility in their realities, the category still considers the context and the physical laws.
4) Image style and location: refer to an image's style and depicted location.
   (a) If an image has multiple styles, select the one that best fits it; if the image does not easily fit into any of the entries, the option "Other" is selected. Select one of the entries: Cartoon, Drawing, Meme, Photograph, or Other.
   (b) Does it show a clear location? Attribute inspired by [33]. Select one of the entries: Indoors: private space, Indoors: public space, Outdoors: private space, Outdoors: public space, or Unclear.
5) Textual Elements: we consider the text's location only, but it could be interesting to add typeface details as well.
   (a) Language: text's original language. Select English, Portuguese, or "N/A" if the image does not contain text.
   (b) Leading Text: text outside of the image that provides context. Select Yes, No, or "N/A" if there is no text.
   (c) Follow-up Text: text that builds off of leading text, providing more context or a punchline. Select Yes, No, or "N/A" if there is no text.
   (d) Integrated Text: any text within the image itself. Select Yes, No, or "N/A" if there is no text.

The Blend Dimension shares two categories with the Emotional Design dimension (both shown with the latter). Categories and attributes are as follows:

1) Resemblance schemes: if there are possible comparisons within an image.
   (a) Resemblance: an umbrella term for visual metaphors, comparison, and personification. Whether a subject in an image appears to imitate something it is not in reality or is compared to something in a way that showcases similarities. For example, an object's shape could naturally resemble that of an animal or human, or it could be artificially manipulated to look like something else, e.g., a cake that looks like a computer. Note: this category refers only to visual comparisons. Select one of the entries: Absent or Present.
   (b) Optical Illusion: if the image contains an element that tricks the viewer's eyes in some way. Select one of the entries: Absent or Present.
   (c) Figure of speech: if the image's textual elements use a "figure of speech", such as a metaphor, personification, or prosopopoeia. Select one of the entries: True, False, Unclear.
2) Outward (ad hoc) Participant/Observer
   (a) Outward Observer: whether the image's observer is assumed to be observing the scene or participating in it in some way. Whether there is an implied observer, who is not explicitly shown in the image but is assumed to exist in order to understand the image's context or meaning (e.g., "POV" memes). Select one of the entries: Absent or Present.
3) Image Context: any relevant contextual information needed to understand the image.
   (a) Context: external factors or circumstances that influence or inform the image's interpretation and meaning. It can include a wide range of concepts, such as cultural

references, historical events, social norms, or even the specific time and place in which the image was created or viewed. No fixed attribute options. E.g., "COVID", a movie's name if knowledge of a certain movie is needed, etc. "N/A" if there are no external contexts necessary for understanding.

(b) Time-situated context: whether the image refers to a specific time frame, such as the pandemic. Select one of the entries: True or False.

4) Call for action: whether it seems to provoke the observer to act.

(a) Call for action: Select one of the entries: True, False, Unclear.

The Emotional Design dimension categories and attributes are as follows:

1) Meaning breakdown: written notes to explain the image and call attention to something particularly unique about the image.

(a) Explanation notes: there are no fixed attribute options, as it should contain short written notes.

2) Emotional-Alignment: this category is shared with the Blend dimension.

(a) Emotional alignment: points to the observer. If the observer is supposed to feel the same/similar emotion as the image's subject (s), then the attribute is considered "aligned". If the intended emotion is different from that of the subject, then the attribute is considered "unaligned". If there is no obvious emotional framing, then is considered "absent". However, if it is ambiguous due to various emotional layers within the image, the attribute is labeled as "Ambiguous". Select one of the entries: Absent, Aligned, Unaligned, Ambiguous.

(b) Irony: whether it conveys irony, either for a humorous effect or not. Select one of the entries: True, False, or Unclear.

3) Humorous Intent and Delivery Method: this category is shared with the Blend dimension.

(a) Intent: whether an image is clearly designed to provide enjoyment or humor. Select one of the entries: True, Neutral, or Opposite (for negative emotions).

(b) Humor Delivery Method: describes how humor is conveyed to the image's observer. Multiple categories can be selected: Absent (the image does not have entertainment/humorous intent), Visual Humor (humor is conveyed using visuals), Textual Humor (is conveyed using text), Pun (humor is conveyed through wordplay), Self-deprecating humor, Other (some form of humor not covered in the previous options).

Prediction of Emergency Department Visits Applying an One Health Approach: Further Investigations

Ismaela Avellino
R&D Researcher
GPI SpA
Trento, Italy
email: ismaela.avellino@gpi.it

Isabella Della Torre
R&D Researcher
GPI SpA
Trento, Italy
email: isabella.dellatorre@gpi.it

Francesca Marinaro
R&D Researcher
GPI SpA
Trento, Italy
email: francesca.marinaro@gpi.it

Andrea Buccoliero
R&D Project Manager
GPI SpA
Trento, Italy
email: andrea.buccoliero@gpi.it

Antonio Colangelo
R&D Director
GPI SpA
Trento, Italy
email: antonio.colangelo@gpi.it

***Abstract*—Proper management of emergency rooms is needed to improve healthcare and patient satisfaction, guiding resource allocation. Predicting access and hospitalisation rates through Machine Learning appears feasible and promising, especially when coupled with air pollution and weather data. This work further investigates, in a more detailed way, a previously presented approach that applied predictive algorithms to data related to Brescia's clinical and environmental data from 2018 to 2022 to predict daily accesses or daily hospitalisations for cardiovascular or respiratory disorders. Starting from the previous work, that analysis was improved and widened to a greater geographical area. The applied algorithms' performances satisfactorily adhere to the actual data, especially when using the Support Vector Machine and Random Forest's models as regressors on daily accesses and respiratory disease-caused hospitalisations. Even if the specific value is not always correctly predicted, generally, the overall trend seems to be rightly forecasted, and performance metrics are rather satisfying. Although additional work could still be encouraged to improve the models' performances, results are rewarding and represent a new point of view on a complex and relevant matter. The real-life application of this One Health approach is now possible and could quite easily be adapted to other areas, too, with the final objective of improving the quality of healthcare and people's quality of life.**

***Keywords-Forecasting; ER accesses; Hospitalisations; Pollution; Weather; One Health; Environmental exposure*.**

I. Introduction

This work is an extension of our previous research presented at the AIHealth 2024 conference that took place in Athens, Greece [1].

It aimed to enable the forecast of the Emergency Department (ED) and Emergency Room (ER) fluxes of patients based on their geographically fixed short-term exposure to pollution agents and weather conditions.

Here, this approach is further investigated and broadened to a larger geographical area, extending the applied methods and reaching a more detailed analysis.

Properly managing ED and ER is crucial to providing functional healthcare and improving patients' satisfaction [2]. It leads to a strong need for accurate prediction of visitor volume and patient admissions to facilitate the planning of resources and staff for the whole hospital.

Multiple researchers have tried to predict access and admission rates based on historical ED data by creating scores or using Deep Learning (DL) or Machine Learning (ML) models (like Recurrent Neural Networks, Logistic Regression, Random Forest or Extreme Gradient Boosting) to forecast daily accesses to the ER [3]–[5], the possibility of a patient's hospital admission after going through the triage [6] or even the risk of death [7]. Results are so encouraging that others continue to look for associations with the surrounding environment.

There is proof that weather affects one's health, especially for people with specific illnesses or healthcare needs. For example, there seems to be a link between the daily temperature and ED admissions for cardiovascular diseases or significant exacerbation of asthma in adults that visit ED [8][9]. Generally speaking, regarding cardiovascular disorders, a worsening of the patient's well-being and cardiac arrests appear to be influenced by temperature and other stressors like humidity and atmospheric pressure [10][11]. Moreover, there is also proof of links between air pollution and specific illnesses. Substances like $PM_{2.5}$, PM_{10}, NO_x, O_3 and SO_2 influence cardiac arrests [12], cardiac arrhythmia [13], cognitive decline in adult population [14], COVID-19 incidence [15], development of chronic kidney disease [16] or Type 2 diabetes [17]. $PM_{2.5}$ and PM_{10} are also linked to hospital admissions for cardiovascular [18] and respiratory diseases [19]. $PM_{2.5}$ levels also seem to be directly associated with increased daily ED visits for ulcerative colitis [20], while solar radiation is inversely associated with inflammatory bowel disease admissions [21]. There also seems to be a correlation between the number of hospitalised asthma patients and both weather (i.e., temperature and humidity) and pollution (i.e., $PM_{2.5}$, PM_{10} and NO_x) [22]. Finally, ML models (i.e., AutoRegressive Integrated Moving Average and Multilayer Perceptron) have also been used to try to predict accesses to the ER by patients affected by infecting respiratory

diseases after being exposed to $PM_{2.5}$ [23].

Some of these investigations are based on long-term exposure to pollution (even 20-years long [14]), while others on a few days or even same day's exposure [15] [18] and some even on both [13].

The amount of days linked to long- or short-term exposure differs for each study and group of researchers, leading to different temporal definitions and freedom of choice when fixing it. For example, when considering only climatic variables, greater exposure can be seven days long [5], meaning that the forecast based on today's data will be projected one week in the future.

Based on these literary pieces of evidence, trying to predict all accesses to the ED or hospitalisation post-triage for specific illnesses, working on climate, pollution and historical accesses time-series belonging to the same area, seems feasible, even if complex.

Indeed, one of the underlying issues of ED visits' prediction is how non-homogeneous and inconstant patients' emergency accesses are. An urgent crisis can suddenly arise without any clear previous sign or from a multitude of variables that are difficult to constantly monitor simultaneously: inpatients' fluxes in ERs and hospitals are ever-changing and subject to the influence of factors like seasons, outbreaks and social conditions [5].

Each year, between 77000 and 80000 patients visit the ER of the largest Brescia hospital [24], and 24% of them get admitted. This is the reason why this ED seemed like the perfect place where to start our attempt at accurately predicting future accesses based on historical and local meteorological and pollution data.

This paper contains a description of the analysed materials and applied methods (i.e., the datasets and the ML approaches applied to them) in Section II, the reached results in Section III, a comment on them in Section IV and a few final remarks in Section V.

II. Materials and Methods

This section describes the study design, analysed datasets (both clinical and environmental data) and applied algorithms.

A. Study Design

This study primarily aims to daily predict the volume of patients going through the ER of a precise hospital in Brescia, Italy.

Forecasting algorithms were designed for ER accesses and hospitalisations from triage for cardiovascular or respiratory diseases.

This retrospective study applies to daily data (clinical and environmental) for a period going from January 1, 2018, to December 31, 2022. A four-year (i.e., 2018–2021) dataset was used to train the forecasting models, while the remaining data were used to test its forecasting capability. The final datasets used to feed the predictive algorithms combine the clinical and the environmental data.

Data collection

The following subsections describe the datasets of interest analysed in this study.

B. Clinical Data

The original clinical dataset was given by a hospital in Brescia to GPI for research purposes.

The dataset contained all anonymous ER access data for the period going from 2018 to 2022. For each access (i.e., a person on a specific day), there were as many rows as the exams the person had undergone; pre-processing was made to have only one row for each ED visit while maintaining the patient's data (like the date of ER visit, their age, sex and zip code of their home address, the list of medical exams they underwent and, in case they went through hospitalisation, their diagnosis as an ICD9-CM code).

The patients came from different cities: most came from the area surrounding the hospital, while others came from other Italian regions or even from abroad. This study's focus was the area for which environmental data had been collected: Brescia. This work presents two different population divisions based on how the Brescia area is geographically identified by the Italian bureaucracy and due to differences in how environmental data were computed to get the best granularity possible. This will be further described in Subsection II-C.

Table I describes the original overall dataset.

TABLE I
Brief description of clinical data.

Year	*Total accesses*	*Median age*	*Male percentage*	*Female percentage*
2018	60176	55	49%	51%
2019	60106	56	49%	51%
2020	47205	58	52%	48%
2021	49571	57	50%	50%
2022	56631	56	51%	49%

In 2018, 12% of patients were below 18 years old, 31% between 19 and 49, 23% between 50 and 69, 34% above 70. In 2019, 11% of patients were below 18 years old, 30% between 19 and 49, 24% between 50 and 69, 35% above 70. In 2020, 9% of patients were below 18 years old, 29% between 19 and 49, 27% between 50 and 69, 35% above 70.
In 2021, 10% of patients were below 18 years old, 30% between 19 and 49, 26% between 50 and 69, 34% above 70. In 2022, 12% of patients were below 18 years old, 29% between 19 and 49, 25% between 50 and 69, 34% above 70. Amongst the most recurrent diagnoses of the hospitalised patients, through all years, were pneumonia and chronic heart failure.

Table II reports the different percentages of ER accesses in the quarters of each analysed year.

The variables included in our final dataset are:

- Categorical information about the date (as described in Table III), from which dummies were computed

TABLE II
DISTRIBUTION OF ER ACCESSES IN THE DIFFERENT YEARS QUARTERS.

Year	*1st quarter*	*2nd quarter*	*3rd quarter*	*4th quarter*
2018	25.7%	25.3%	24.2%	24.8%
2019	26.7%	24.5%	23.7%	25.1%
2020	29.7%	23.4%	24.7%	22.2%
2021	22.7%	24.8%	25.8%	26.7%
2022	23.1%	25.7%	25.0%	26.2%

- Daily number of accesses to the ER or hospitalisations coming from it, limited to those patients coming either from just the city of Brescia or also from its entire province
- The rolling mean of the number of accesses or hospitalisations, applying a seven-day window for calculation.

TABLE III
DESCRIPTION OF CALENDRICAL INFORMATION.

Calendrical variable	*Definition*
Day of the week	Monday, Tuesday, [...], Saturday, Sunday
Day of the month	1, 2, 3, 4, [...], 28, 29, 30, 31
Month	January, February, [...], November, December
Year	2018, 2019, 2020, 2021, 2022

The subdivisions in different pathological groups were done by selecting the correct hospitalisations through the ICD9-CM codes reported as the primary diagnosis for their access.

Table IV describes the dataset restricted to the city of Brescia.

TABLE IV
BRIEF DESCRIPTION OF CLINICAL DATA (CITY OF BRESCIA).

Year	*Total accesses*	*Median age*	*Male percentage*	*Female percentage*
2018	10389	56	46%	54%
2019	10963	58	47%	53%
2020	9835	61	50%	50%
2021	11082	60	49%	51%
2022	12597	60	49%	51%

In 2018, 11% of patients were below 18 years old, 30% between 19 and 49, 24% between 50 and 69, 35% above 70.
In 2019, 10% of patients were below 18 years old, 29% between 19 and 49, 24% between 50 and 69, 37% above 70.
In 2020, 8% of patients were below 18 years old, 27% between 19 and 49, 27% between 50 and 69, 38% above 70.
In 2021, 9% of patients were below 18 years old, 28% between 19 and 49, 25% between 50 and 69, 38% above 70.
In 2022, 11% of patients were below 18 years old, 27% between 19 and 49, 23% between 50 and 69, 39% above 70.

Table V describes the dataset widened to Brescia's province.

In 2018, 12% of patients were below 18 years old, 31% between 19 and 49, 24% between 50 and 69, 33% above 70.
In 2019, 11% of patients were below 18 years old, 30% between 19 and 49, 24% between 50 and 69, 35% above 70.
In 2020, 9% of patients were below 18 years old, 28% between 19 and 49, 27% between 50 and 69, 36% above 70.
In 2021, 10% of patients were below 18 years old, 29% between 19 and 49, 26% between 50 and 69, 35% above 70.
In 2022, 12% of patients were below 18 years old, 28% between 19 and 49, 25% between 50 and 69, 35% above 70.

TABLE V
BRIEF DESCRIPTION OF CLINICAL DATA (BRESCIA'S PROVINCE).

Year	*Total accesses*	*Median age*	*Male percentage*	*Female percentage*
2018	53378	55	48%	52%
2019	53678	57	49%	51%
2020	43445	59	49%	51%
2021	46386	58	50%	50%
2022	52518	57	50%	50%

A little contextualisation of the clinical dataset: it is fundamental to note that the area around Brescia suffered substantially from the outbreak of the COVID-19 pandemic, and the number of cases affected by coronavirus pneumonia far exceeds the occurrences of any other diagnosis during 2020.

It is possible to observe from Table I and Table V, and this is something already reported in other studies [25] [26], that the number of accesses to ER significantly decreased from 2019 to 2020: this is explainable because Italy was in a strict lockdown for several months that year. Hence, it was less likely, for example, for car accidents to happen or for people wearing masks to get the flu.

Note that, regarding our data of interest, while this trend is observable for both the general accesses and those from Brescia's province, it is not valid for the patients from the city itself.

C. Environmental Data

The environmental data have been supplied by the startup Hypermeteo [27] under GPI's specific request to match the spatio-temporal dimension of the already-at-disposal clinical dataset.

The environmental data for the city of Brescia are defined per day and zip (the Italian CAP) code, guaranteeing spatial-temporal precision. On the contrary, the province area is defined by ISTAT codes, while the corresponding zip code was also reported, aggregating them.

Two different codes describe Italian municipalities: CAP and ISTAT. The first is a postal code, while the other links to the homonymous Italian statistics authority [28].

These environmental data were obtained employing a mathematical model with a resolution of $10km$x$10km$, corrected through normalisation and down-scaling, and applied to data by Lombardia's Regional Environmental Protection Agency (ARPA [29]) weather stations.

While the model was built for the entire Lombardia region, environmental data were extracted for the province of Brescia only, and our study was divided into two phases.

In fact, at first, only data from the city of Brescia itself were analysed, and the results of this approach were reported

in our previous publication [1]. Now, we have re-approached the same city data but also widened our analysis to the entire province.

When working with the sole city of Brescia, its 15 zip codes were differentiated both in the clinical and environmental data and were all linked to only 1 ISTAT code. This, unfortunately, was not the case for the province data.

In this sense, the city of Brescia and its province are differently identified. While every municipality is linked to one and only one ISTAT code, Brescia's city is further defined into 15 different CAP codes, where one CAP code can define multiple of its province's municipalities.

Since the ISTAT code can be linked to many different CAP codes and the environmental province data was defined based on the former, if we wanted to link the clinical dataset to the environmental one, we had to find a way to reduce the latter to one row per zip code.

For this reason, for the same zip code, we computed the mean of each environmental variable for each day, enabling the later merge between this dataset and the clinical one.

Apart from the different identifying geographical codes, the rest of the datasets are precisely the same for both approaches, describing the same variables.

Specifically, the reported environmental variables are:

- Temperature (min and max values) (T_{min}, T_{max} [$^{\circ}C$])
- Humidity (min and max percentage values) (RH_{min}, RH_{max} [$\%$])
- Precipitations (Prec [mm])
- PM_{10} and $PM_{2.5}$ [$\mu g/m^3$]
- NO_x, SO_2, NMVOC and O_3 [$\mu g/m^3$]
- Total solar irradiance (SSW_{tot}) [Wh/m^2].

For each variable, safety ranges, provided along with the dataset, were considered in order to give a label (i.e., zero or one) to each value to indicate if a value could be safe or not, respectively. Depending on the variable, either lower or upper bounds were considered, as reported in Table VI. These safety ranges have been chosen with Hypermeteo based on institutional guidelines [30].

TABLE VI
SAFETY RANGES FOR ENVIRONMENTAL VARIABLES.

Environmental variable	Lower and Upper Bounds	
	Min value	*Max value*
NO_x	-	25 $\mu g/m^3$
$PM_{2.5}$	-	15 $\mu g/m^3$
PM_{10}	-	45 $\mu g/m^3$
SO_2	-	40 $\mu g/m^3$
NMVOC	-	1000 $\mu g/m^3$
O_3	-	100 $\mu g/m^3$
T_{min}	-10 $^{\circ}C$	-
T_{max}	-	35 $^{\circ}C$
RH_{min}	15%	-
RH_{max}	-	95%
Prec	-	10 mm
SSW_{tot}	-	8500 Wh/m^2

Subsequently, we computed the number of occurrences in which the data were out of range for the city and province datasets. Occurrences are a single day of the five years considered per single zip code.

In the city of Brescia, in around the 71% of occurrences NO_x was out of range, it was the 60% of cases for $PM_{2.5}$, 20% for PM_{10}, 17% for RH_{max}, 8% for the precipitations, 7.5% for O_3, 2% for T_{max}, 0.5% for SSW_{tot}, 0.3% for RH_{min} and 0 cases out of range for NMVOC, SO_2 and T_{min}.

In its province, in around the 44% of occurrences NO_x was out of range, it was the 46% of cases for $PM_{2.5}$, 13% for PM_{10}, 21% for RH_{max}, 8% for the precipitations, 5% for O_3, 0.8% for T_{max}, 0.4% for SSW_{tot}, 0.4% for RH_{min}, 0.4 for T_{min} and 0 cases out of range for NMVOC and SO_2.

The issue of having multiple rows of data for the same date (i.e., one row for each zip code) has been handled similarly as in a project [31] found during our bibliographic research: each environmental variable has been labelled with the zip code it is referred to, and it is used as a column with daily values, thus grouping all data belonging to the same date on one row.

Again, a clarification on the context: the area surrounding Brescia is densely inhabited and industrialised, resulting in one of the most polluted areas in Europe [32].

PREDICTIVE ALGORITHMS

The following subsections describe the different predictive algorithms applied to the analysed datasets: Random Forest, Artificial Neural Network, Support Vector Machine and AutoRegressive Integrated Moving Average.

D. Random Forest

In order to predict future ER accesses or hospitalisations, based on our clinical and environmental data, the first algorithm to be applied was the same as the one used in the previous study [1].

A Random Forest (RF) approach was implemented in Python with the application of the open-source library Scikit-learn [33]. This model was chosen based on an article [34] that applied it to a temperature prediction problem: the analogy with our dataset highlighted this approach as a fascinating candidate for this type of analysis.

RFs apply sequential splits to the data such that the separation is maximised in regards to a homogeneity criterion, resulting in a combination of tree predictors so that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [4].

The random forest algorithm picks N random records from the dataset and builds a decision tree based on them, repeating the process for the selected number of trees. The topic has been tackled as a regression problem as we have considered the target variable (i.e., daily accesses) as continuous.

The main goal of our modelling approach was to create an algorithm that improved the error compared with the average baseline one (Average Baseline Error, ABE), which we computed as the mean absolute difference between the actual values and the rolling mean. We considered the latter the most basic prediction to be produced since it simply uses the

rolling mean of the target variable calculated for the previous seven days as the predicted future value.

To find the best parameters to set the RF model to, we applied a Python optimisation library called Optuna [35] that, through multiple trials, finds the values that minimise or maximise a specific metric of interest. In our case, we opted to minimise the MAE.

Trying to improve performances (both in terms of metrics and computational time), we applied Optuna to obtain the optimal parameters (n_estimators, max_depth, min_samples_split, min_samples_leaf) for the RF. For each case study, we ran multiple trials to reach their best combination.

The results that are reported in Subsection III-A are based on different combinations of the datasets, as we applied the same model on both the city of Brescia's and the entirety of its province's whole datasets, and then, again, for both of them, on different data combinations and only on cardiovascular or respiratory disorders data.

Since we also worked on the entire province, we widened the application of the same logic used in our previous study [1] to its data.

To predict the future values of interest, we, again, applied a temporal lag to the datasets, but, this time, only one day (and not five too). This means that the observed data from the previous day is used to predict the volume of patient accesses or hospitalisations on the subsequent one.

The different analyses that were carried out, trying to improve the model's performance and potentially spot specific influential variables, can be divided into seven macro cases:

1) To enable further discussion over the preciseness of our daily accesses' predictions and validate our datasets' composition, we decided to deepen our analysis on what we considered to be our baseline.
   Thus, in addition to computing the ABE, we also decided to apply a model constructed using the same number of trees as applied during the previous study [1] to each spatial dataset (city and province) reduced to only contain the rolling mean and calendrical information, thus without any environmental feature.
2) The RF algorithm was applied to the initial preprocessed accesses' dataset made up of patients from the city of Brescia:
   a) at first, the applied model was created using the same number of trees as applied during the previous study [1],
   b) then, the best model was searched, and the best combination of its parameters was found in order to produce the best achievable prediction,
   c) finally, this last model was applied to only the two most important (as computed by the best model itself) features.
3) The RF algorithm was applied to the initial preprocessed hospitalisations dataset made up of patients from the city of Brescia and whose main diagnosis was a cardiovascular disease:
   a) at first, the best model was searched and found by optimising its parameters,
   b) then, it was applied to only the two most important (as computed by the best model itself) features.
4) The RF algorithm was applied to the initial preprocessed hospitalisations dataset made up of patients from the city of Brescia and whose main diagnosis was a respiratory disease:
   a) at first, the best model was searched and found by optimising its parameters,
   b) then, it was applied to only the two most important (as computed by the best model itself) features.
5) The RF algorithm was applied to the initial preprocessed accesses' dataset made up of patients from the entire province of Brescia:
   a) at first, the applied model was created using the same number of trees as applied during the previous study [1],
   b) then, the best model was searched, and the best combination of its parameters was found in order to produce the best achievable prediction,
   c) later, trying to improve the metrics, we casually divided the first four years (2018-2021) into train (80%) and test (20%) that we input into a trial for the best model and then used it to predict our last available year (2022, our usual year of test). We have done so as it looked like using a casual division gave better metrics' values,
   d) then, a model with the same configuration as the best one was applied to only the two most important (as computed by the best model itself) features,
   e) finally, the study on the most important features was reapplied, not to create a new RF model but rather to study which environmental features have the most influence on the prediction when discarding the rolling mean or both the rolling mean and calendrical information about the different days.
6) The RF algorithm was applied to the initial preprocessed hospitalisations' dataset made up of patients from the entire province of Brescia and whose main diagnosis was a cardiovascular disease:
   a) at first, the best model was searched and found by optimising its parameters,
   b) then, again, a study on which environmental features have the most influence on the prediction (thus probably also on the hospitalisations) was conducted.
7) The RF algorithm was applied to the initial preprocessed hospitalisations' dataset made up of patients from the entire province of Brescia and whose main diagnosis was a respiratory disease:
   a) at first, the best model was searched and found by optimising its parameters,

b) then, again, a study on which environmental features have the most influence on the prediction (thus probably also on the hospitalisations) was conducted.

To evaluate the performances of our models, we applied different metrics.

First, we computed the ABE to be considered as the value to be improved.

Then, we also computed the mean and standard deviation (that will be reported as dispersion in Section III) of both the actual and predicted values.

Here, the equations for the other metrics applied to evaluate the models' performances are reported. They were Mean Absolute Error (MAE, 1), Mean Absolute Percentage Error (MAPE, 2), Symmetric Mean Absolute Percentage Error (SMAPE, 3), and R^2 score (4) [36]:

$$MAE = \sum_{i=1}^{N} |\hat{y}_i - y_i| \tag{1}$$

$$MAPE = \frac{100}{N} \sum_{i=1}^{N} \frac{y_i - \hat{y}_i}{y_i} \tag{2}$$

$$SMAPE = \frac{100}{N} \sum_{i=1}^{N} \frac{|\hat{y}_i - y_i|}{(|y_i| + |\hat{y}_i|)/2} \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{4}$$

where N is the test sample size, y is the actual values' vector, $\hat{y}_i$ is the predicted values' one and $\bar{y}$ is the mean of the actual test values.

Since the applied algorithm is a regressor, the usual Accuracy equation cannot be used. It was replaced with a value we called Acc* that we computed using the MAPE subtracted from 100% as reported in Equation 5. MAPE is sometimes called Forecast Error Percentage, so it seemed fitting to create such a metric.

$$Acc^* = 100 - MAPE \tag{5}$$

We also plotted the comparison graphs between the actual and predicted values for the daily accesses and hospitalisations. We also plotted their smoothed version to appreciate the preciseness of the forecast more, as data were noisy. The applied smoothing filter was Savitzky-Golay, with a temporal window length of 31 days and a polynomial order of 2.

E. Artificial Neural Network

As in the previous study [1], trying to improve the results given by the algorithm described in Subsection II-D, other algorithms were applied to hospitalisation data.

Specifically, the first was an Artificial Neural Network (ANN) [37] designed in Python. This model was only used on hospitalisation data linked to cardiovascular or respiratory diseases of patients from both the city and the province of Brescia.

Since ANN is a distance-dependent model, trying to achieve the best performance possible, we applied scaling on the data through a specific library [38].

The used model was a 2-layer shallow neural network, and an optimisation algorithm was, again, applied to search for the best parameters possible.

The selected metrics to evaluate the performances were MAE (1) and SMAPE (3).

Once more, we plotted the comparison graphs between the actual and predicted values for the daily cardiovascular hospitalisations and their smoothed version computed by applying the same filter described in Subsection II-D.

F. Support Vector Machine

Further trying to improve the prediction of hospitalisations, a Support Vector Machine (SVM) [39] was implemented.

It is a supervised ML algorithm that, in this case, we used for regression and applied in Python through its homonymous library [40]. Its main aim is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes, guaranteeing a margin between the closest points of different classes to be the maximum possible.

When a Support Vector is applied to solve regression problems, its produced model depends only on a subset of the training data because the cost function ignores samples whose prediction is close to their target.

Our implementation applied a Linear Support Vector Regressor fine-tuned through the Python application of a Scikit-learn library called GridSearchCV [41].

This model was applied only to the hospitalisation data for both spatial (city and province) datasets.

The applied metrics to evaluate the performances were MAE (1) and R^2 score (4).

We plotted the comparison graphs between the actual and predicted values for the daily hospitalisations and their smoothed version computed by applying the same filter described in Subsection II-D.

G. AutoRegressive Integrated Moving Average

In the previous study [1], trying to improve the results given by the algorithm described in Subsection II-D, a specific ML model for multivariate time-series prediction was applied to the hospitalisation data: an AutoRegressive Integrated Moving Average (ARIMA) model [42].

It is a popular algorithm used in time series analysis and forecast.

For the previous analysis, we applied the auto-ARIMA process [43] in Python, while, this time, we applied another library that enabled the guided research of the best parameters: statsmodels' ARIMA function [44].

The basic idea of the ARIMA model is to use a particular mathematical algorithm to describe the random time series of the data and then predict the future values based on the

past and present values through a so-called autoregression. An ARIMA (p, d, q) model can be described through Equation 6:

$$(1 - \sum_{i=1}^{p} \varphi_i L^i)(1 - L)^d X_t = (1 + \sum_{i=1}^{q} \theta_i L^i)\varepsilon_t \quad (6)$$

where L represents the lag operator, p represents the number of autoregressive terms, q represents the number of moving average terms, d represents the degree of differencing, and ϕ, θ and ϵ are relevant parameters.

Since the achieved results were, again, not promising, we are not going to report all of them, but just an interesting aspect about the city's respiratory hospitalisations' time-series prediction from this model that highlights a peculiar characteristic of the actual data.

The reported results come from the application of the ARIMA model on hospitalisations due to respiratory diseases of patients from the city of Brescia.

III. Results

This section reports the obtained results from the various predictive algorithms.

Even if the algorithms have been fed with different datasets, they always include only data related to patients whose home address' zip code is either inside the city of Brescia or its entire province, based on their objective as described in Subsections II-D, II-E, II-F and II-G.

As already declared, the presented results are performance metrics' values or plots.

The second ones show the curves representing the daily predicted values (always plotted in magenta) versus the actual values for the testing year (i.e., 2022), plotted in different colours based on the predictive algorithm they come from.

Note that when metrics could not be computed due to data sparsity, they were not reported for that specific case study.

A. Random Forest

This subsection presents the results of the RF application to our datasets of interest.

City of Brescia

Please note that the results reported for the datasets constituted by accesses and hospitalisations of patients from the zip codes of Brescia (the same dataset analysed in the previous study [1]) have been improved and newly computed.

1) Daily accesses' baseline: As previously anticipated, to further evaluate the goodness of our RF models for the daily accesses' predictions, we analysed datasets reduced to only the rolling mean and calendrical information.

The first metric to be computed was the ABE, as it was considered to be the value to improve. It was equal to 4.92.

Here are the results for the 1000 trees model:

- MAE = 4.83
- R^2 score = 0.21
- Acc* = 85.63%
- MAPE = 14.37%
- SMAPE = 6.9%
- Mean of actual accesses = 34.51
- Dispersion of actual accesses = 6.69.

2) Daily accesses: The following values are the metrics computed for the model created using the original number of trees (i.e., 1000):

- MAE = 4.75
- MAPE = 14.37%
- SMAPE = 6.9%
- Acc* = 85.63%
- R^2 score = 0.21
- Mean of predicted accesses = 34.20
- Dispersion of predicted accesses = 3.22.

The following reported values are the metrics computed for the best model coming from the optimisation (i.e., 423 trees):

- MAE = 4.63
- MAPE = 13.91%
- SMAPE = 6.8%
- Acc* = 86.09%
- R^2 Score = 0.24
- Mean of predicted accesses = 33.89
- Dispersion of predicted accesses = 2.78.

Figures 1 and 2 plot the actual (in blue) and predicted (in magenta) accesses and their smoothed version, respectively.

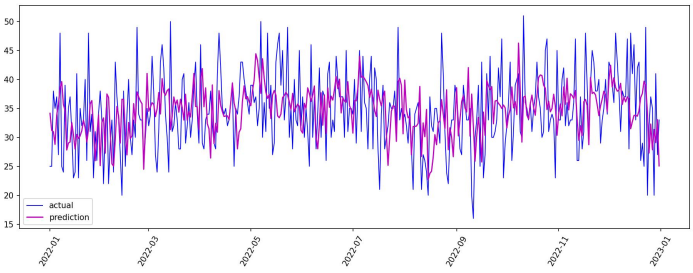


Figure 1. Random Forest's predicted (as computed by the best model) and actual values of daily ER accesses for Brescia.

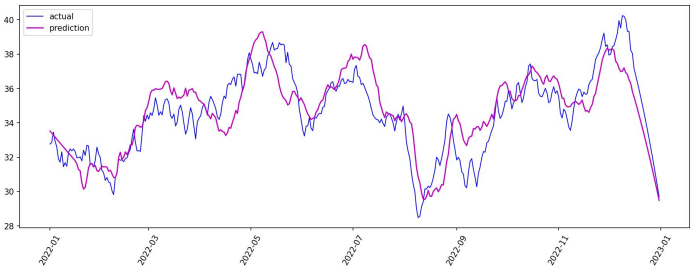


Figure 2. Random Forest's smoothed predicted (as computed by the best model) and actual values of daily ER accesses for Brescia.

The following metrics are the ones computed from the model whose input were just the two most important features (resulting from the best model):

- MAE = 5.56
- Acc* = 82.82%.

These features were the rolling mean and the day of the month.

3) Daily hospitalisations for cardiovascular diseases: We computed the ABE to use as the value to be improved, and it was equal to 0.49.

The following reported values are the metrics computed for the best model coming from the optimisation (i.e., 855 trees):

- MAE = 0.51
- SMAPE = 77.86%
- R^2 Score: 0.08
- Mean of actual hospitalisations: 0.45
- Mean of predicted hospitalisations = 0.48
- Dispersion of actual hospitalisations = 0.68
- Dispersion of predicted hospitalisations = 0.31.

Figures 3 and 4 plot the actual (in blue) and predicted (in magenta) cardiovascular hospitalisations and their smoothed version, respectively.

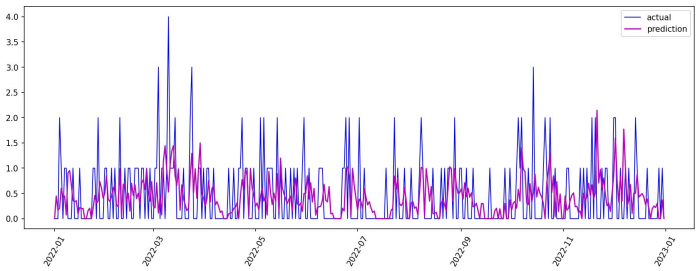


Figure 3. Random Forest's predicted (as computed by the best model) and actual values of daily hospitalisations for cardiovascular diseases for Brescia.

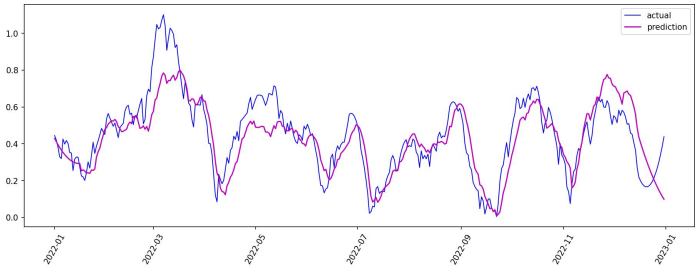


Figure 4. Random Forest's smoothed predicted (as computed by the best model) and actual values of daily hospitalisations for cardiovascular diseases for Brescia.

The value of MAE computed from the model whose input were just the two most important features (resulting from the best model) was 0.49. These features were the rolling mean and the day-of-the-month information.

4) Daily hospitalisations for respiratory diseases: We computed the ABE to use as the value to be improved: it was equal to 1.05.

The following reported values are the metrics computed for the best model coming from the optimisation (i.e., 991 trees):

- MAE = 1.05
- SMAPE = 33.7%
- R^2 Score = 0.22
- Mean of actual hospitalisations = 2.02
- Mean of predicted hospitalisations = 1.99
- Dispersion of actual hospitalisations = 1.48
- Dispersion of predicted hospitalisations = 0.79.

Figures 5 and 6 plot the actual (in blue) and predicted (in magenta) respiratory hospitalisations and their smoothed version, respectively.

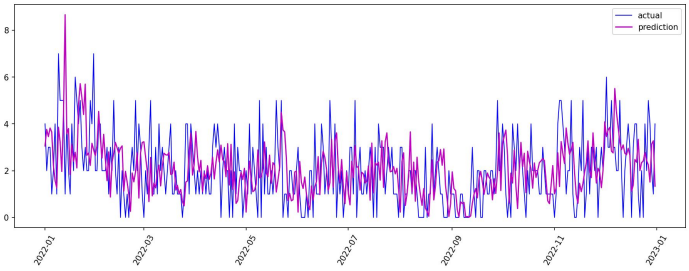


Figure 5. Random Forest's predicted (as computed by the best model) and actual values of daily hospitalisations for respiratory diseases for Brescia.

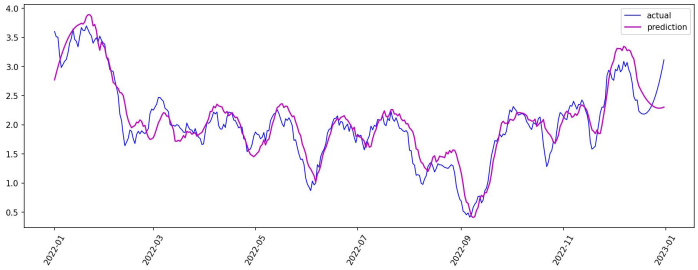


Figure 6. Random Forest's smoothed predicted (as computed by the best model) and actual values of daily hospitalisations for respiratory diseases for Brescia.

The value of MAE computed from the model whose input were just the two most important features (resulting from the best model) was 1.23. These features were the rolling mean and the day-of-the-month information.

Brescia's province

This section presents the analysis conducted on data of patients from the entire province of Brescia, which was never considered in the previous study [1].

5) Daily accesses' baseline: The computed value of ABE, the metric to be improved, was 12.79.

Again, here are reported the reached results for the 1000 trees model analysis of the baseline dataset:

- MAE = 10.49
- R^2 score = 0.51
- Acc* = 92.58%
- MAPE = 7.42%
- SMAPE = 3.69%
- Mean of actual accesses = 143.88
- Dispersion of actual accesses = 18.23.

6) Daily accesses: The following values are the metrics computed for the model created using the original number of trees (i.e., 1000):

- MAE = 9.81

- MAPE = 6.95%
- SMAPE = 3.45%
- Acc* = 93.05%
- R² Score = 0.55
- Mean of predicted accesses = 142.72
- Dispersion of predicted accesses = 12.56.

The following reported values are the metrics computed for the best model coming from the optimisation (i.e., 396 trees):

- MAE = 9.58
- MAPE = 6.81%
- SMAPE = 3.36%
- Acc* = 93.19%
- R² Score = 0.57
- Mean of predicted accesses = 143.15
- Dispersion of predicted accesses = 12.30.

Figures 7 and 8 plot the actual (in blue) and predicted (in magenta) accesses and their smoothed version, respectively.

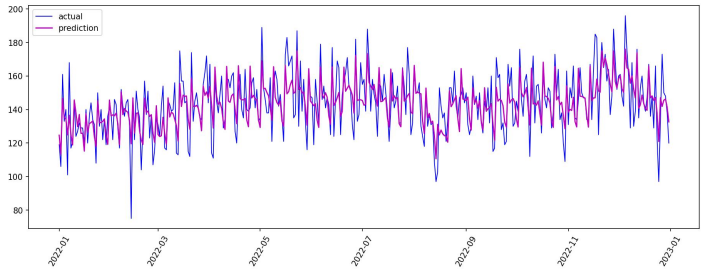


Figure 7. Random Forest's predicted (as computed by the best model) and actual values of daily ER accesses for the whole province of Brescia.

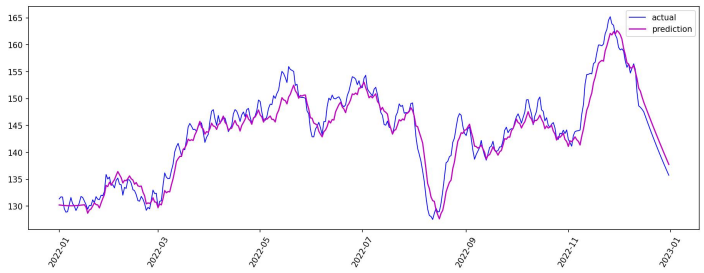


Figure 8. Random Forest's smoothed predicted (as computed by the best model) and actual values of daily ER accesses for the whole province of Brescia.

Computing a new best model (i.e., 940 trees), we tried a new approach. We divided the first four years of the dataset into the train and test portions casually rather than chronologically, and the obtained metrics were:

- MAE = 9.63
- R² Score = 0.74.

If we then used this same model to predict, as usual, the accesses for 2022 (as they represented completely new data for the algorithm), the metrics were:

- MAE = 9.84
- R² Score = 0.54.

The following metrics are those computed from the model whose input were just the two most important features (resulting from the best model). These features were the rolling mean and the Monday label.

- MAE = 12.78
- Acc* = 90.77%.

We were also interested to see which environmental variables were the most influential on the daily accesses, so we computed the best model and the feature importance for a dataset extracted from the original one without the rolling mean and on another where we removed the information about the days, too.

In the first case, the most important environmental variable was NO_x, while in the second case, the most important variables were NO_x, PM_{10}, RH_{max}, $PM_{2.5}$ and T_{min}.

7) Daily hospitalisations for cardiovascular diseases: The computed value of ABE was 0.81.

The following reported values are the metrics computed for the best model coming from the optimisation (i.e., 1112 trees):

- MAE = 0.87
- SMAPE = 39.8%
- R² Score = 0.06
- Mean of actual hospitalisations = 1.37
- Mean of predicted hospitalisations = 1.38
- Dispersion of actual hospitalisations = 1.13
- Dispersion of predicted hospitalisations = 0.49.

Figures 9 and 10 plot the actual (in blue) and predicted (in magenta) cardiovascular hospitalisations and their smoothed version, respectively.

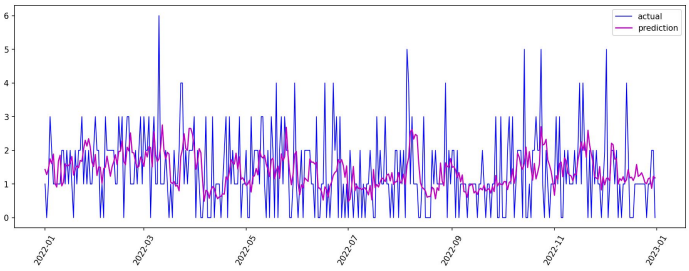


Figure 9. Random Forest's predicted (as computed by the best model) and actual values of daily hospitalisations for cardiovascular diseases for the whole province of Brescia.

We were also interested to see which environmental variables were the most influential on the daily cardiovascular hospitalisations, so we computed the best model and the feature importance for a dataset extracted from the original one without the rolling mean and on another where we removed the information about the days, too.

In the first case, the most important environmental variables were NO_x, RH_{max}, T_{min} and PM_{10}; in the second case they were the same, with the addition of $PM_{2.5}$.

8) Daily hospitalisations for respiratory diseases: The computed value of ABE was 1.95.

The following reported values are the metrics computed for the best model coming from the optimisation (i.e., 105 trees):

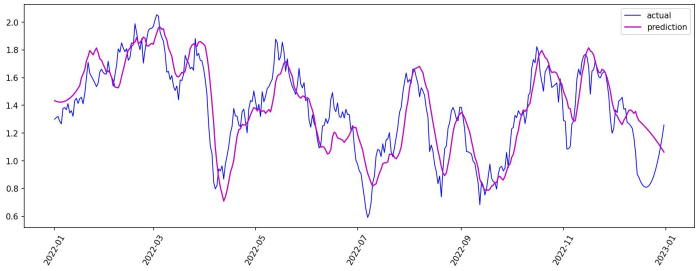


Figure 10. Random Forest's smoothed predicted (as computed by the best model) and actual values of daily hospitalisations for cardiovascular diseases for the whole province of Brescia.

- MAE = 1.96
- SMAPE = 14.3%
- R² Score = 0.45
- Mean of actual hospitalisations = 7.60
- Mean of predicted hospitalisations= 7.53
- Dispersion of actual hospitalisations = 3.35
- Dispersion of predicted hospitalisations = 2.37.

Figures 11 and 12 plot the actual (in blue) and predicted (in magenta) respiratory hospitalisations and their smoothed version, respectively.

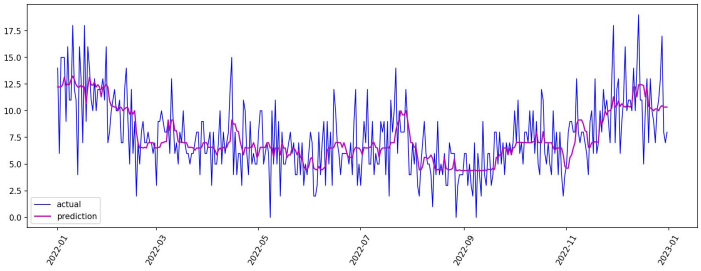


Figure 11. Random Forest's predicted (as computed by the best model) and actual values of daily hospitalisations for respiratory diseases for the whole province of Brescia.

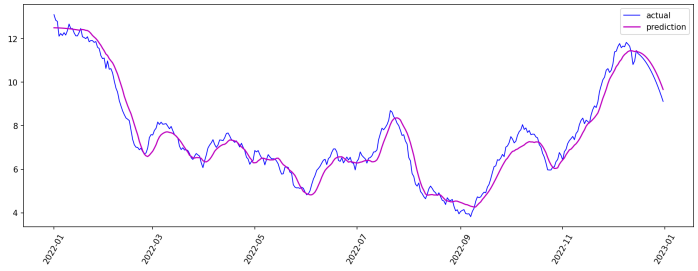


Figure 12. Random Forest's smoothed predicted (as computed by the best model) and actual values of daily hospitalisations for respiratory diseases for the whole province of Brescia.

We were also interested to see which environmental variables were the most influential on the daily respiratory hospitalisations, so we computed the best model and the feature importance for a dataset extracted from the original one without the rolling mean and on another where we removed the information about the days, too.

In the first case, the most important environmental variables were NO_x and T_{min}, while in the second case, they were T_{min}, $PM_{2.5}$, NO_x, PM_{10}, O_3 and RH_{max}.

B. Artificial Neural Network

Here will be reported the metrics and plots resulting from the application (described in Subsection II-E) of a shallow 2-layer ANN to the hospitalisations caused by cardiovascular or respiratory disorders for patients coming both from only the city of Brescia and those from its entire province too.

Again, both numerical results of metrics and graphs are reported.

City of Brescia

1) Daily hospitalisations for cardiovascular diseases: The ABE to be improved was equal to 0.49, and the computed MAE for the ANN applied to the hospitalisations for cardiovascular diseases for Brescia was equal to 0.53. The SMAPE was 78.51%.

Figures 13 and 14 plot the actual (in green) and predicted (in magenta) Brescia's cardiovascular hospitalisations and their smoothed version, respectively.

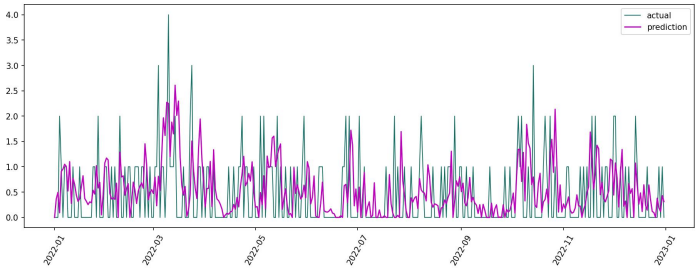


Figure 13. Artificial Neural Network's predicted and actual values of daily hospitalisations for cardiovascular diseases for Brescia.

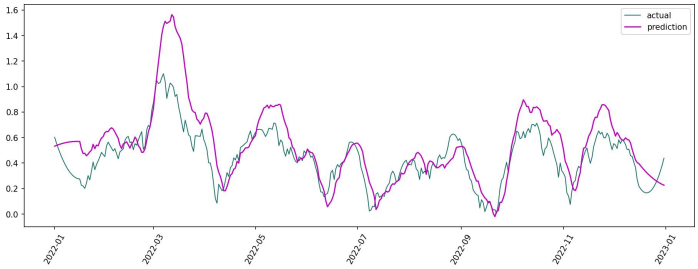


Figure 14. Artificial Neural Network's smoothed predicted and actual values of daily hospitalisations for cardiovascular diseases for Brescia.

2) Daily hospitalisations for respiratory diseases: The ABE to be improved was equal to 1.05, and the computed MAE for the ANN applied to the hospitalisations for respiratory diseases for Brescia was equal to 1.19. The SMAPE was 39.46%.

Figures 15 and 16 plot the actual (in green) and predicted (in magenta) Brescia's respiratory hospitalisations and their smoothed version, respectively.

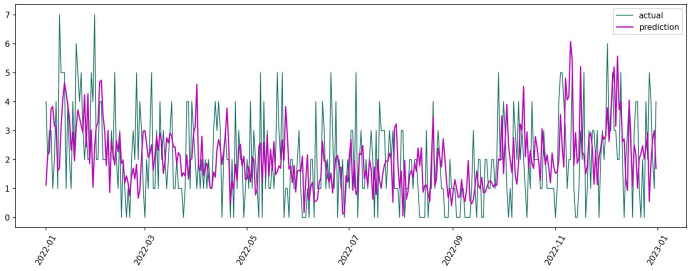


Figure 15. Artificial Neural Network's predicted and actual values of daily hospitalisations for respiratory diseases for Brescia.

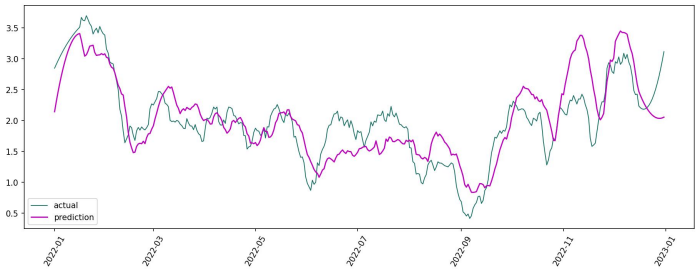


Figure 16. Artificial Neural Network's smoothed predicted and actual values of daily hospitalisations for respiratory diseases for Brescia.

BRESCIA'S PROVINCE

3) Daily hospitalisations for cardiovascular diseases: The ABE to be improved was equal to 0.81, and the computed MAE for the ANN applied to the hospitalisations for cardiovascular diseases for the province of Brescia was equal to 1.17. The SMAPE was 47.79%.

Figures 17 and 18 plot the actual (in green) and predicted (in magenta) cardiovascular hospitalisations and their smoothed version, respectively.

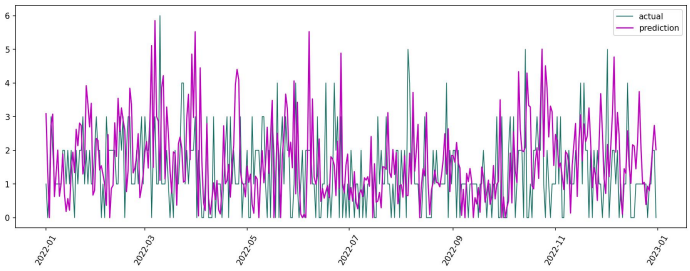


Figure 17. Artificial Neural Network's predicted and actual values of daily hospitalisations for cardiovascular diseases for the whole province of Brescia.

4) Daily hospitalisations for respiratory diseases: The ABE to be improved was equal to 1.95, and the computed MAE for the ANN applied to the hospitalisations for respiratory diseases for the province of Brescia was equal to 3.34. The SMAPE was 20.00%.

Figures 19 and 20 plot the actual (in green) and predicted (in magenta) respiratory hospitalisations and their smoothed version, respectively.

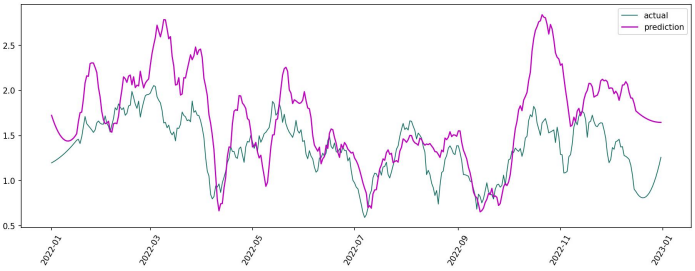


Figure 18. Artificial Neural Network's smoothed predicted and actual values of daily hospitalisations for cardiovascular diseases for the whole province of Brescia.

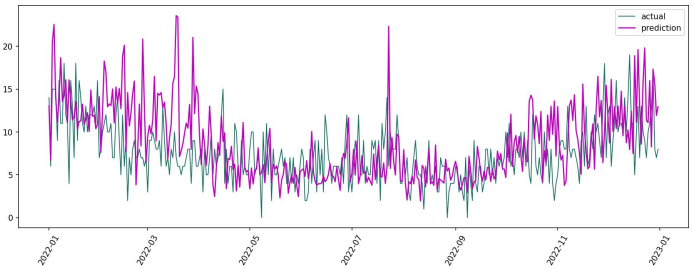


Figure 19. Artificial Neural Network's predicted and actual values of daily hospitalisations for respiratory diseases for the whole province of Brescia.

C. Support Vector Regression

Here are the results of the approach described in Subsection II-F to analyse and improve the predictions of daily hospitalisations for both cardiovascular and respiratory diseases for the city and province of Brescia.

As always, both numerical results of metrics and graphs are reported.

CITY OF BRESCIA

1) Daily hospitalisations for cardiovascular diseases: The ABE to be improved was equal to 0.49, and the computed MAE for the SVR applied to the hospitalisations for cardiovascular diseases from Brescia was 0.50. The R^2 Score was 0.09.

Figures 21 and 22 plot the actual (in grey) and predicted (in magenta) cardiovascular hospitalisations and their smoothed version, respectively.

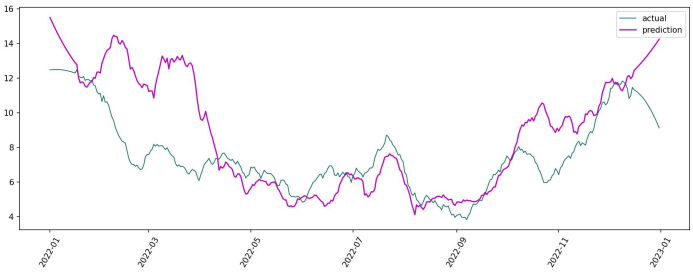


Figure 20. Artificial Neural Network's smoothed predicted and actual values of daily hospitalisations for respiratory diseases for the whole province of Brescia.

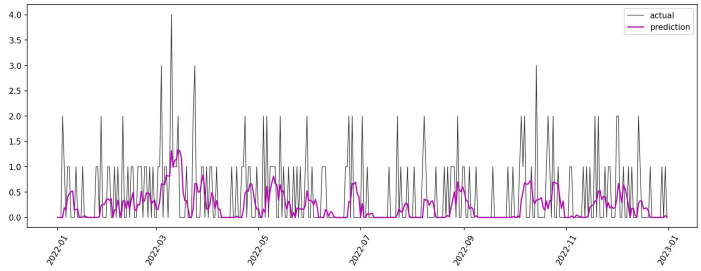


Figure 21. Support Vector Machine's predicted and actual values of daily hospitalisations for cardiovascular diseases for Brescia.

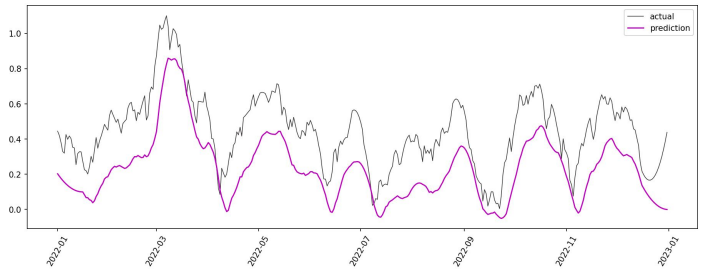


Figure 22. Support Vector Machine's smoothed predicted and actual values of daily hospitalisations for cardiovascular diseases for Brescia.

2) Daily hospitalisations for respiratory diseases: The ABE to be improved was equal to 1.05, and the computed MAE for the SVR applied to the hospitalisations for respiratory diseases from Brescia was 1.04. The R^2 Score was 0.39.

Figures 23 and 24 plot the actual (in grey) and predicted (in magenta) Brescia's respiratory hospitalisations and their smoothed version, respectively.

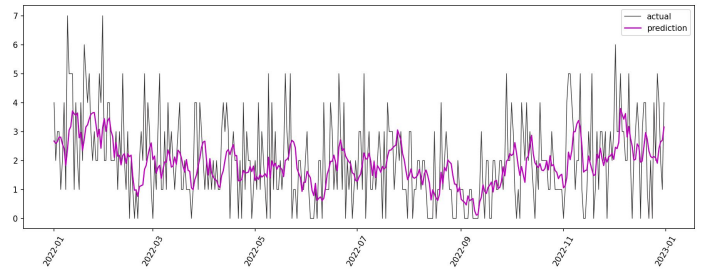


Figure 23. Support Vector Machine's predicted and actual values of daily hospitalisations for respiratory diseases for Brescia.

Brescia's province

3) Daily hospitalisations for cardiovascular diseases: The ABE to be improved was equal to 0.81, and the computed MAE for the SVR applied to the hospitalisations for cardiovascular diseases from Brescia was 0.83. The R^2 Score was 0.12.

Figures 25 and 26 plot the actual (in grey) and predicted (in magenta) cardiovascular hospitalisations and their smoothed version, respectively.

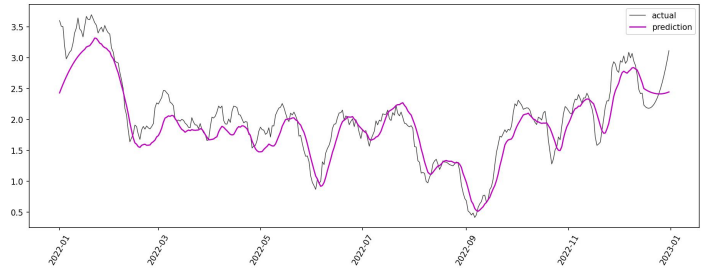


Figure 24. Support Vector Machine's smoothed predicted and actual values of daily hospitalisations for respiratory diseases for Brescia.

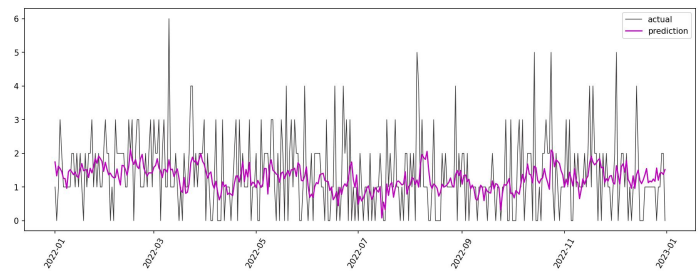


Figure 25. Support Vector Machine's predicted and actual values of daily hospitalisations for cardiovascular diseases for the whole province of Brescia.

4) Daily hospitalisations for respiratory diseases: The ABE to be improved was equal to 1.95, and the computed MAE for the SVR applied to the hospitalisations for respiratory diseases from Brescia was 1.94. The R^2 Score was 0.66.

Figures 27 and 28 plot the actual (in grey) and predicted (in magenta) respiratory hospitalisations and their smoothed version, respectively.

D. ARIMA

As already noted in Subsection II-G, here will be reported only one striking aspect of the daily hospitalisations for respiratory diseases of patients from Brescia.

Figure 29 plots the actual values coming from the train (2018-2021) portion of the respiratory hospitalisation dataset for the city, while Figure 30 the predictions (in magenta) of the 2022's test values (in brown).

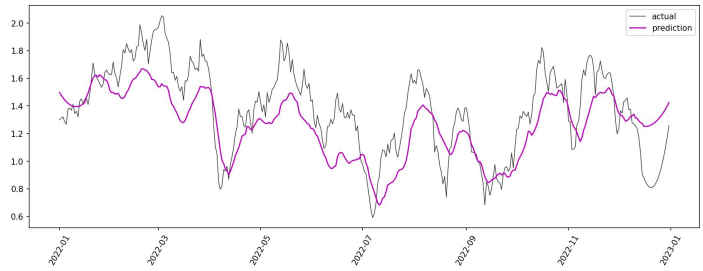


Figure 26. Support Vector Machine's smoothed predicted and actual values of daily hospitalisations for cardiovascular diseases for the whole province of Brescia.

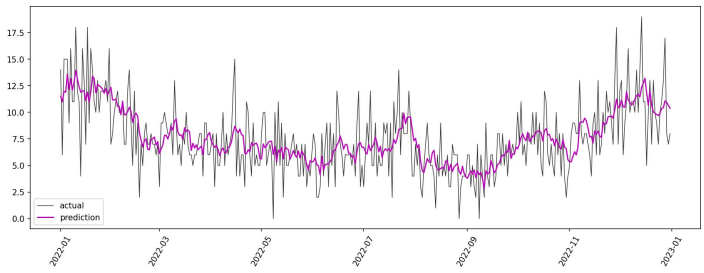


Figure 27. Support Vector Machine's predicted and actual values of daily hospitalisations for respiratory diseases for the whole province of Brescia.

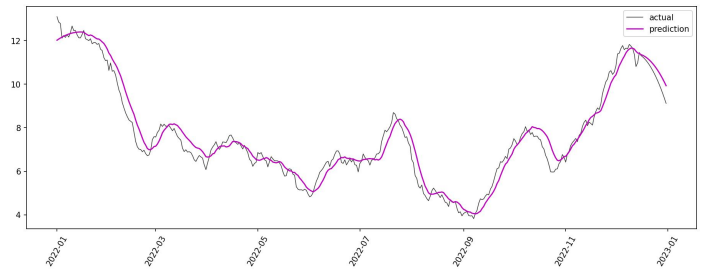


Figure 28. Support Vector Machine's smoothed predicted and actual values of daily hospitalisations for respiratory diseases for the whole province of Brescia.

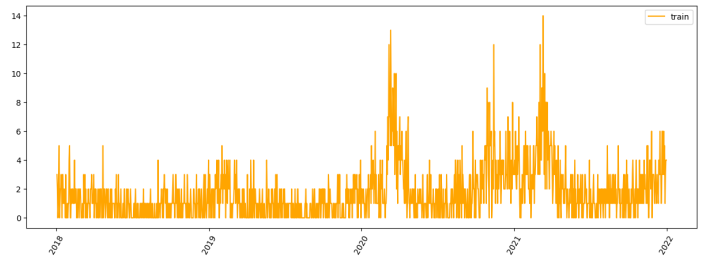


Figure 29. Actual daily hospitalisations caused by respiratory diseases for patients from Brescia from 2018 to 2021.

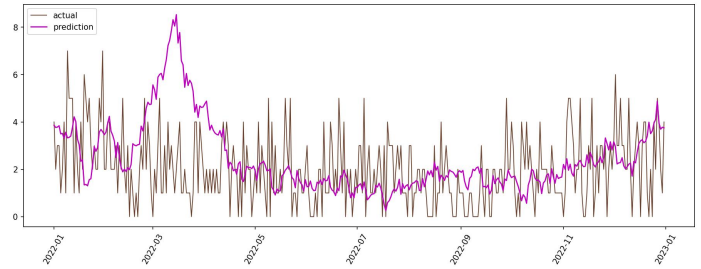


Figure 30. ARIMA's predicted and actual values of daily hospitalisations for respiratory diseases for Brescia.

IV. Discussion

The results, obtained applying the different predictive algorithms, reported in Section III will now be discussed.

A. Random Forest

The following are evaluations and comments on the results reported in Subsection III-A.

City of Brescia

1) Daily accesses: The results reported in Subsubsection III-A2, referring to the daily accesses of patients coming only from the city of Brescia, will now be discussed.

The error to improve (i.e., ABE) was 4.92. The achieved results for the baseline model represent the goodness of a poor prediction and can be used to evaluate if and how adding environmental data can improve the forecast.

The RF with the same number of estimators as the previous paper [1] already had better performances as its MAE was lower (i.e., 4.75).

The Acc* of this prediction, as computed from MAPE, was an appreciable 85.63%, and SMAPE was 6.9%. Since SMAPE accounts for the relative and balanced difference between predicted and actual values, such a low score indicates that the model predicts rather well.

Trying different approaches to evaluate how close the predictions were to the actual values, we computed their mean and dispersion. The obtained results confirm that, even though the dispersion is not as large, the mean is quite close, indicating that the general trend has been rightly forecasted.

The R^2 score (i.e., 0.21) was not high, but it was positive and, considering the complexity of the analysed scenario, can be deemed acceptable.

In an attempt to improve the performances further, the best model was computed over several trials, and the achieved MAE was even lower and equal to 4.63.

Even though the mean and dispersion of the actual and predicted values were not as good, our main objective was to minimise the MAE, and this model succeeded. Moreover, the Acc* was higher (i.e., 86.09%), so the MAPE was minor, and the SMAPE value was slightly too.

Unfortunately, the R^2 score was still not especially solid, as it was only 0.24.

The plots visually confirm the predictor's considerably satisfactory ability to follow the general trend.

Comparing these results with the analysis done on the dataset without environmental features, we can see how adding this type of information results in better metrics, thus producing a more precise forecast.

The metrics resulting from the best model applied to only the two most important features (none of which was environmental) were not as satisfying as the ones from the whole dataset, further proving that adding climate and pollution data influences the prediction positively.

2) Daily cardiovascular hospitalisations: The results reported in Subsubsection III-A3, referring to the daily hospitalisations due to cardiovascular diseases of patients coming only from Brescia, will now be discussed.

Since the model was, in this case, applied to sparse data, the Acc* could not be computed.

The error to improve was 0.49. The smaller MAE value was 0.51, obtained by applying the best model with 855 trees, and still worse than the ABE.

The R^2 score was too small (and the smallest yet) to be adequate, as it was too close to 0. Confirming this consideration, SMAPE, reaching 77.86%, was significantly greater than the others.

It appears clear that the approach needed to be changed to reach a better forecast of these accesses.

Regarding the mean and dispersion of the actual and predicted values, the former only had a 0.03 difference, while the latter had a 0.37 one. It means the forecast has a close but slightly narrower point cloud of accesses.

Still, the plots, especially the smoothed one, show that the model forecasts the general trend with acceptable precision.

The MAE resulting from the application only to the two most important features (none of which was environmental) was smaller than the one coming from the whole dataset but still not lower than ABE, so it cannot be considered successful.

3) Daily respiratory hospitalisations: The results reported in Subsubsection III-A4, referring to the daily hospitalisations due to respiratory diseases of patients coming only from Brescia, will now be discussed.

Since the model was, again, applied to sparse data, the Acc* could not be computed.

The error to improve was 1.05, and the best model (with 991 trees) achieved a MAE value equal to it. Since the value has remained equal and not lowered, the model cannot yet be considered satisfying.

R^2 score and SMAPE were not satisfying either: the approach needed to be changed to better forecast these accesses.

Regarding the mean and dispersion of the actual and predicted values, the former only had a 0.03 difference, and the latter a 0.69 one.

Even though the metrics are not ideal, the plots, especially the smoothed one, show that the general trend has been quite rightly forecasted.

The MAE resulting from the application on only the two most important features (none of which was environmental) was even higher than the one coming from the whole dataset, which was merely equal to the ABE. It further proves that adding climate and pollution data influences the predictor performances positively.

Brescia's province

4) Daily accesses: The results reported in Subsubsection III-A6, referring to the daily accesses of patients from the entire province of Brescia, will now be discussed.

The error to improve was 12.79. The RF with the same number of estimators as the previous paper already had better performances as its MAE was lower (i.e., 9.81).

The Acc* of this prediction, as computed from MAPE, was a satisfying 93.05%.

Trying different approaches to evaluate how close the predictions were to the actual values, we computed their mean and dispersion. The obtained results confirm that, even though the dispersion is not as large, the mean is quite close, indicating that the general trend has been rightly forecasted.

Even if the R^2 score was only 0.55, the model seems valid, and SMAPE was only 3.45%, so the prediction's errors are reasonably negligible, resulting in the best values reached for these metrics yet, thus the most precise model.

In trying to improve these performances further, the best model was computed over several trials, and the achieved metrics were, in fact, even better.

The MAE, equal to 9.58, was even lower, and the Acc* of the prediction (i.e., 93.19%) was slightly higher.

Regarding the mean and dispersion of the actual and predicted values, the first one was even closer, while the dispersion marginally worsened. Still, the general trend has been rightly forecasted.

Even if the R^2 score was only 0.57, it is still the best achieved one, considering all previous case studies, while SMAPE was the smallest one, as it was equal to only 3.36%.

The plots visually confirm the predictors' ability to follow the general trend.

When trying to divide the train and test datasets casually instead of chronologically, the metrics appeared to be better as we reached, through its own best model, an R^2 Score as high as 0.74.

For this reason, we tried further using this different approach, but we also had to test it on future chronologically presented data, as that is how future input data would look.

Unfortunately, though, when tested on 2022 data, the model performance returned to values closer to the ones from the initial chronological division. So, we decided to discard this plan and revert to the original one.

The metrics resulting from applying the best model on only the two most important features (none of which was environmental) were not as satisfying as the ones from the whole dataset, proving further that adding climate and pollution data influences the prediction positively.

Regarding the influence of environmental variables on ER accesses, it is interesting to note that low temperatures and humidity rates hold this much of an impact, as some polluting substances do. The fact that minimum temperature and pollution substances appear together is unsurprising since heating systems release pollutants like $PM_{2.5}$ and PM_{10}.

5) Daily cardiovascular hospitalisations: The results reported in Subsubsection III-A7, referring to the daily hospitalisations due to cardiovascular diseases of patients from the entire province of Brescia, will now be discussed.

Since, for this case study, the model was applied again to sparse data, the Acc* could not be computed.

The error to improve was 0.81. The lower MAE value was 0.87, through the application of the best model having 1112 trees, nevertheless worse than the ABE.

The R^2 score (i.e., 0.06) was even smaller than the city (reported in Subsubsection III-A3) one, even if the SMAPE (i.e., 39.8%) was minor. This worse R^2 score could be due to the added sparsity of data from adding patients that follow the same noisy general trend (way different than the whole accesses' one).

Clearly, the approach needed to be changed to forecast these accesses better.

Regarding the mean and dispersion of the actual and predicted values, the former only had a 0.01 difference, while the latter had a 0.64 one. It means that the forecast has a narrower point cloud of accesses.

Still, the plots, especially the smoothed one, show that the model forecasts the general trend with acceptable precision.

Regarding the influence of environmental variables on cardiovascular hospitalisations coming from triage, it is interesting to note that humidity and temperature have such an impact, along with some polluting substances.

6) Daily respiratory hospitalisations: The results reported in Subsubsection III-A8, referring to the daily hospitalisations due to respiratory diseases of patients from the entire province of Brescia, will now be discussed.

Since the model was, again, applied to sparse data, the Acc^* could not be computed.

The error to improve was 1.95. The best model (with 105 trees) achieved a MAE value of 1.96, still slightly higher than the ABE, an R^2 score of 0.45, and a SMAPE of 14.3%. Even though these last two values were better than the ones reached for the city's respiratory hospitalisations and less unsatisfactory than the MAE, the approach still needed to be changed to achieve a valid forecast.

Regarding the mean and dispersion of the actual and predicted values, the former only had a 0.07 difference, and the latter a 0.98 one.

Still, the plots, especially the smoothed one, show that the general trend has been satisfyingly forecasted.

Regarding the influence of environmental variables on respiratory hospitalisations coming from triage, it is interesting to note that minimum temperature has such an impact, along with more polluting substances (compared with the other case studies). It was expected because of evidence that respiratory disorders flares link to air pollution [9] [19].

B. Artificial Neural Network

The following are evaluations and comments on the results reported in Subsection III-B.

City of Brescia

1) Daily cardiovascular hospitalisations: The results reported in Subsubsection III-B1, referring to the daily hospitalisations of patients affected by cardiovascular diseases coming from Brescia, will now be discussed.

The ABE to improve was 0.49, but, unfortunately, the network's MAE (i.e., 0.53) was higher, even more than the RF one.

SMAPE was 78.51%, again, worse than the RF one.

It further proved that the best predictive algorithm approach for this analysis was yet to be found. The plots do not appear remarkably different from the RF ones, but, nevertheless, not as good as them.

2) Daily respiratory hospitalisations: The results reported in Subsubsection III-B2, referring to the daily hospitalisations of patients affected by respiratory diseases coming from Brescia, will now be discussed.

The ABE to improve was 1.05, but, unfortunately, the network's MAE (i.e., 1.19) was higher and even more than the RF one.

SMAPE was 39.46%, again, worse than the RF one.

It further proved that, even if the plots do not appear tragically different from the actual values, the best predictive algorithm approach for hospitalisations was yet to be found.

Brescia's province

3) Daily cardiovascular hospitalisations: The results reported in Subsubsection III-B3, referring to the daily hospitalisations of patients affected by cardiovascular diseases coming from the entire province of Brescia, will now be discussed.

The ABE to improve was 0.81, but the network's MAE (i.e., 1.17) was still unsatisfactory and worse than the RF one. Same for SMAPE as it was higher.

It deeply proved that the best predictive algorithm approach for this analysis was yet to be found, as the plots appear clearly different and worse than the RF ones.

4) Daily respiratory hospitalisations: The results reported in Subsubsection III-B4, referring to the daily hospitalisations of patients affected by respiratory diseases from the entire province of Brescia, will now be discussed.

The ABE to improve was 1.94, and the network's MAE (i.e., 3.34) was absolutely unsatisfactory and way worse than the RF one. SMAPE was surprisingly small as it was equal to 20%, but still higher than the RF one.

It further proved that the best predictive algorithm approach for this hospitalisation analysis was yet to be found, as the plots appear to diverge from the actual values significantly.

C. Support Vector Machine

The following are evaluations and comments on the results reported in Subsection III-C.

Since ANN did not improve as hoped, we approached another algorithm, finally obtaining better results for one of the two hospitalisations' groupings.

City of Brescia

1) Daily cardiovascular hospitalisations: The results reported in Subsubsection III-C1, referring to the daily hospitalisations due to cardiovascular diseases of patients coming only from Brescia, will now be discussed.

The ABE was 0.49, and the reached MAE was 0.50. Even if it is not better than the baseline error, it is still slightly an improvement, compared to the RF error.

Instead, the R^2 score was dramatically lower because it was 0.09.

By visually analysing the plots, it can be commented that the SVR prediction underestimates the daily hospitalisations.

2) Daily respiratory hospitalisations: The results reported in Subsubsection III-C2, referring to the daily hospitalisations due to respiratory diseases of patients coming only from Brescia, will now be discussed.

The ABE was 1.05, and the reached MAE was 1.04. It represents a case study where the application of a different model did, indeed, improve performances.

Further proving this point, the RF's R^2 score was only 0.22, while SVR's was 0.39.

The plots appear way more adherent, too, resulting in a satisfying forecast of a notably complex application.

BRESCIA'S PROVINCE

3) Daily cardiovascular hospitalisations: The results reported in Subsubsection III-C3, referring to the daily hospitalisations due to cardiovascular diseases of patients coming only from the entire province of Brescia, will now be discussed.

The ABE was 0.81, and the reached MAE was 0.83. Even if it is not better than the baseline error, it is still an improvement compared to the RF one.

The same goes for the R^2 score since it even doubled.

Compared with the RF plots, these appear less adherent to actual data, and they are slightly underestimating.

4) Daily respiratory hospitalisations: The results reported in Subsubsection III-C4, referring to the daily hospitalisations due to respiratory diseases of patients coming only from the whole province of Brescia, will now be discussed.

The ABE was 1.95, and the reached MAE was 1.94. Again, this represents another time when applying a predictive model improved performances. Even the best RF model did not obtain a MAE value smaller than ABE.

Further proving this point, the R^2 score was a striking 0.66, the highest value of this metric we reached in any trial, as we discarded the non-chronological approach.

The plots appear way more adherent, too, especially the smoothed one.

It resulted in the best forecast of all, even though we must highlight that we have not applied SVR to daily accesses as we had already found valid models, so we do not know which results would have come out of that.

D. ARIMA

The following are evaluations and comments on the results reported in Subsection III-D.

Based on the previous findings [1], we already knew that ARIMA was not the ideal model to improve the performances of our forecast, but we still decided to run it to see if we could find any aspect of interest.

Since ARIMA is a time-series-based analysis, trend fluxes heavily influence it: this resulted in a peculiar prediction graph for respiratory diseases-caused hospitalisations of patients coming from Brescia, as it predicted a phantom positive peak around March.

As we investigated the reason for that, we found its explanation in the observable trend of the previous years' actual data: in fact, March 2020 and 2021 saw a surge in hospitalisations due to respiratory disorders as more patients contracted COVID-19.

The main drawback of time series models is that they rely only upon the forecasted variable without comprehending and looking for the concealed causes of its behaviour. Still, they can represent a suitable approach when dealing with real-life daily chronological data.

V. CONCLUSION AND FUTURE WORK

When analysing metrics and graphs from the different models, we can appreciate how, in the end, for both the city of Brescia and its province, we could manage to validly predict daily accesses and hospitalisations due to respiratory diseases.

The same cannot be said for cardiovascular hospitalisations, plausibly due to the high sparsity of these data, meaning that further research needs to be undertaken. Note that the number of hospitalisations for specific pathologies is limited to a few people every day and, sometimes, even none, and this is particularly noticeable for cardiovascular disorders.

Still, the main objective of this work, which was to upgrade and deepen the previously reported analysis [1], was generally reached. Even the worst result, coming from the analysis of cardiovascular hospitalisations of patients, still represents an improvement from the previous study, and the latter's findings have been validated.

Focusing on comparing the different predictive algorithms, we can state that, for these specific datasets, SVR seems to be the best one, followed by RF. ANN, instead, results in performances closer to the ones of ARIMA.

Visually analysing the plots, our best forecasts of daily accesses and respiratory hospitalisations appear to adhere quite well to the actual data, and their metrics are quite satisfying, too.

In fact, generally speaking, even if the specific values are not always correctly predicted, the overall trend seems to be rightly followed, and peak values (like surges in accesses or hospitalisations) are captured.

Another significant result to highlight is the confirmation of how adding environmental data can improve the prediction.

When we tried to apply the same models to reduced versions of the datasets that only contained calendrical information or, instead, discarded it, we generally achieved better performances.

Based on these observations, this work represents a coherent deep-dive that further analyses the previous approach.

The prediction of ER accesses and hospitalisations from a specific geographical area through the analysis of clinical and environmental data is feasible.

The previous promising results have been confirmed and improved, even if this method's application on cardiovascular hospitalisations could still benefit from further investigation.

Nevertheless, we cannot generalise the results since we obtained them by analysing a period majorly made up of

COVID-19-ridden years and a limited geographical area. Thus, we can only use them to comment on this specific frame.

The performances could dramatically differ if the analogous pre-processing and the same models were applied to other contexts or just even on a longer and more stable period.

In summary, our hypothesis of enabling forecasting of ER volumes by combining historical clinical, weather and pollution data, linked by a detailed geographical indication, has been proven to be suitable and also given more than encouraging results.

Although additional work could still be encouraged to improve the achieved performances, this represents a new point of view on such a complex and poignant matter.

The real-life application of this approach is now possible, and its adaptation to other areas appears simple, even if we cannot predict how accurate that forecast would be.

To conclude, future developments of this work will widen to other areas, with the hope of moving to ever-growing datasets, and additional algorithm testing will be conducted to improve the best-achieved predictions further.

Nevertheless, any additional attempt will gather supplementary valuable insight on this topic and shed light on how our surrounding environment influences human health.

This One Health approach may offset a new way of managing ER worldwide, enabling the monitoring of entire populations and geographical areas, with the final objective of improving the quality of healthcare and people's quality of life.

REFERENCES

[1] I. Della Torre, I. Avellino, F. Marinaro, A. Buccoliero, and A. Colangelo, "Predictive analytics for Emergency Department visits based on local short-term pollution and weather exposure", AIHealth 2024, The First International Conference on AI-Health. ThinkMind, pp. 29-34, 2024.

[2] J. D. Sonis and B. A. White, "Optimizing patient experience in the emergency department", Emergency Medicine Clinics, vol. 38, no. 3, pp. 705–713, 2020.

[3] Z. Qiao et al., "Using machine learning approaches for emergency room visit prediction based on electronic health record data", Building continents of knowledge in Oceans of data: The future of co-created eHealth. IOS Press, pp. 111–115, 2018.

[4] Y. M. Chiu, J. Courteau, I. Dufour, A. Vanasse, and C. Hudon, "Machine learning to improve frequent emergency department use prediction: a retrospective cohort study", Scientific Reports, vol. 13, no. 1, p. 1981, 2023.

[5] C. Peláez-Rodríguez, R. Torres-López, J. Pérez-Aracil, N. López-Laguna, S. Sánchez-Rodríguez, and S. Salcedo-Sanz, "An explainable machine learning approach for hospital emergency department visits forecasting using continuous training and multi-model regression" Computer Methods and Programs in Biomedicine, vol. 245, 2024.

[6] A. Cameron, K. Rodgers, A. Ireland, R. Jamdar, and G. A. McKay, "A simple tool to predict admission at the time of triage", Emergency Medicine Journal, vol. 32, no. 3, pp. 174–179, 2015.

[7] R. Sánchez-Salmerón et al., "Machine learning methods applied to triage in emergency services: A systematic review", International Emergency Nursing, vol. 60, 2022.

[8] W. Zhu et al., "The effect and prediction of diurnal temperature range in high altitude area on outpatient and emergency room admissions for cardiovascular diseases", International Archives of Occupational and Environmental Health, vol. 94, no. 8, pp. 1783–1795, 2021.

[9] T. Abe et al., "The relationship of short-term air pollution and weather to ED visits for asthma in Japan", The American journal of emergency medicine, vol. 27, no. 2, pp. 153–159, 2009.

[10] D. Martinaitiene and N. Raskauskiene, "Weather-related subjective well-being in patients with coronary artery disease", International Journal of Biometeorology, vol. 65, pp. 1299–1312, 2021.

[11] M. Hensel et al., "Association between weather-related factors and cardiac arrest of presumed cardiac etiology: a prospective observational study based on out-of-hospital care data", Prehospital Emergency Care, vol. 22, no. 3, pp. 345–352, 2018.

[12] S. Kojima et al., "Fine particulate matter and out-of-hospital cardiac arrest of respiratory origin", European Respiratory Journal, vol. 57, no. 6, p. 2004299, 2021.

[13] M. A. Shahrbaf, M. A. Akbarzadeh, M. Tabary, and I. Khaheshi, "Air pollution and cardiac arrhythmias: a comprehensive review", Current Problems in Cardiology, vol. 46, no. 3, p. 100649, 2021.

[14] J. M. Delgado-Saborit et al., "A critical review of the epidemiological evidence of effects of air pollution on dementia, cognitive function and cognitive decline in adult population", Science of the Total Environment, vol. 757, p. 143734, 2021.

[15] S.-T. Zang et al., "Ambient air pollution and COVID-19 risk: evidence from 35 observational studies", Environmental research, vol. 204, p. 112065, 2022.

[16] M.-Y. Wu, W.-C. Lo, C.-T. Chao, M.-S. Wu, and C.-K. Chiang, "Association between air pollutants and development of chronic kidney disease: a systematic review and meta-analysis", Science of the Total Environment, vol. 706, p. 135522, 2020.

[17] Y. Li, L. Xu, Z. Shan, W. Teng, and C. Han, "Association between air pollution and type 2 diabetes: an updated review of the literature", Therapeutic Advances in Endocrinology and Metabolism, vol. 10, pp. 1-15, 2019.

[18] R. D. Brook et al., "Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association", Circulation, vol. 121, no. 21, pp. 2331–2378, 2010.

[19] F. Dominici et al., "Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases", Jama, vol. 295, no. 10, pp. 1127–1134, 2006.

[20] R. Duan et al., "Association between short-term exposure to fine particulate pollution and outpatient visits for ulcerative colitis in Beijing, China: A time–series study", Ecotoxicology and Environmental Safety, vol. 214, p. 112-116, 2021.

[21] F. Jaime et al., "Solar radiation is inversely associated with inflammatory bowel disease admissions", Scandinavian journal of gastroenterology, vol. 52, no. 6-7, pp. 730–737, 2017.

[22] C.-L. Chan et al., "A survey of ambulatory-treated asthma and correlation with weather and air pollution conditions within Taiwan during 2001–2010", Journal of asthma, vol. 56, no. 8, pp. 799–807, 2019.

[23] J. Lu et al., "Feasibility of machine learning methods for predicting hospital emergency room visits for respiratory diseases", Environmental Science and Pollution Research, vol. 28, pp. 29701–29709, 2021.

[24] https://civile.asst-spedalicivili.it/servizi/unitaoperative/unitaoperative_fase02.aspx?ID=586 [Retrieved online: November 2024].

[25] F. Tartari, A. Guglielmo, F. Fuligni, and A. Pileri, "Changes in emergency service access after spread of COVID-19 across Italy", Journal of the European Academy of Dermatology and Venereology, vol. 34, no. 8, p. e350, 2020.

[26] T. Ferrari, C. Zengarini, F. Bardazzi, and A. Pileri, "In-depth, single-centre, analysis of changes in emergency service access after the spread of COVID-19 across Italy", Clinical and Experimental Dermatology, vol. 46, no. 8, pp. 1588–1589, 2021.

[27] https://www.hypermeteo.com/ [Retrieved online: November 2024].

[28] https://www.istat.it/en/classification/codes-of-italian-municipalities-provinces-and-regions/ [Retrieved online: November 2024].

[29] https://www.arpalombardia.it/ [Retrieved online: November 2024].

[30] https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health [Retrieved online: November 2024].

[31] https://www.kaggle.com/datasets/nicholasjhana/energy-consumption-generation-prices-and-weather [Retrieved online: November 2024].

[32] S. Khomenko et al. "Premature mortality due to air pollution in European cities: a health impact assessment", The Lancet Planetary Health, vol. 5, no. 3, pp. e121–e134, 2021.

[33] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html [Retrieved online: November 2024].

[34] https://towardsdatascience.com/random-forest-in-python-24d0893d51c0 [Retrieved online: November 2024].

[35] https://optuna.org/ Retrieved online: November 2024].

[36] G. Vishwakarma, A. Sonpal, and J. Hachmann, “Metrics for benchmarking and uncertainty quantification: Quality, applicability, and best practices for machine learning in chemistry”, Trends in Chemistry, vol. 3, no. 2, pp. 155-156, 2021.

[37] J. Zou, Y. Han, and S. S. So, “Overview of artificial neural networks”, Artificial neural networks: methods and applications, pp. 14-22, 2009.

[38] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html [Retrieved online: November 2024].

[39] 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, July-Aug. 1998.

[40] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html [Retrieved online: November 2024].

[41] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html [Retrieved online: November 2024].

[42] R. H., Shumway, and D. S., Stoffer, “ARIMA models”, Time series analysis and its applications: with R examples, pp 75-163, 2017.

[43] https://alkaline-ml.com/pmdarima/modules/generated/pmdarima.arima.auto_arima.html [Retrieved online: November 2024].

[44] https://www.statsmodels.org/stable/generated/statsmodels.tsa.arima.model.ARIMA.html [Retrieved online: November 2024].

Understanding Practice Stages for a Proficient Piano Player to Complete a Piece: Focusing on the Interplay Between Conscious and Unconscious Processes

Katsuko T. Nakahira
Nagaoka University of Technology
Nagaoka, Niigata, Japan
Email: katsuko@vos.nagaokaut.ac.jp

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

Makoto Toyota
T-Method
Chiba, Japan
Email: pubmtoyota@mac.com

***Abstract*— Research into instrumental music performance has garnered significant attention, particularly regarding the intricate interplay of perceptual-cognitive-motor interactions, knowledge application, and the cognitive representation of musical structure. Understanding these dynamics holds promise for enhancing instruction and aiding learners in their journey towards mastering instrumental performance and practice. However, grasping the learning process necessitates more than just comprehending the individual cognitive mechanisms at play; it requires a holistic approach that considers the cognitive architecture enabling the integration of these processes. In this paper, based on the MHP/RT framework proposed by Kitajima and CCE research method which based on the MHP/RT principles, we attempt to understand the process of proficiency in music performance by proficient piano players as a brain model based on the coordination of perception, cognition, and movement, and the concept of Two Mind. Initially, we modeled the cognitive process of piano performance proficiency, and ethnographically described the process of proficiency in music performance for selected elite monitors. The descriptions are analyzed and compared with the model of cognitive processes and actual behaviors in performance proficiency. The description of which perspectives can/cannot be interpreted by the model based on MHP/RT were considered. Finally, a series of piano playing exercises and lessons are analyzed from the perspectives of the Two Minds process, and the knowledge system (implicit/explicit) utilized. Through the analysis, the relationship between acquired knowledge and cognitive ability and Two Minds is considered. The findings suggest that the proficiency process of instrumental music performance exhibits a kind of phase transition. It involves not only a gradual shift from prolonged, System 2-driven mechanical training towards an intuitive, System 1-driven unconscious expression but also deviations from this pattern. Therefore, it is imperative for players to thoroughly comprehend their perception of the entire piece (System 2) while also fostering a sense of ease and naturalness in performance akin to unconscious expression (System 1) for the listener.**

Keywords*— *Proficient Piano Player; Cognitive Process; Two Minds; MHP/RT; Ethnological Study.

I. Introduction

This paper is based on the previous work originally presented in COGNITIVE2024 [1]. A review of MHP/RT and the fundamentals needed to understand the process of training and performance of proficient piano player in preparation for competition were added in Section II.

Instrumental performance has attracted attention as a result of the interaction of perceptual/cognitive and motor abilities. Numerous studies focus on the process of instrumental performance proficiency. The goal of this study is to understand the proficiency process of instrumental performance, which has the possibility of providing better instruction to a performance learner.

Palmer [2] summarizes empirical research on instrumental performance in terms of conceptual interpretation formation, control over motor actions, interpretive transfer as perception, and structural disambiguation. Lehmann and Ericsson [3] focus on the development of instrumental performance skills at the level reached by high school students and amateurs. In their study, they posit that the method of practice is particularly important in improving the level of instrumental performance. A study that focused on the subjectivity factor of instrumental performance practice itself, shares a different perspective; Araújo [4] conducted an online questionnaire survey of self-regulated practice behaviors pertaining to advanced musicians, from which he indicates that practice organization, personal resources, and external resources are important factors. For understanding proficiency in instrumental performance, Chaffin et al. [5][6] applied the protocol analysis method, investigating the characteristics of a concert pianist's performance of a piece of music, in addition to the characteristics of the music. They categorized elements of the instrumental performance in three basic dimensions (fingering, high difficulty, and familiarity with the note form), four interpretive dimensions (phrasing, dynamics, tempo, and pedal), and three expressive dimensions (basic, interpretative, and expressive). Through the categorization process, a possibility of the existence of image for desired representation of the music from the beginning, so-called a "*big picture*", was found.

Focusing on *how to practice* instrumental music performance, as Palmer [2] mentioned, an individual's cognitive representation of musical structure is important especially from the perspectives of specific errors and knowledge utilization in instrumental music performance. To understand this, it is not sufficient to understand the cognitive mechanisms for individual perceptual, cognitive, and motor processes, but research from the perspective of cognitive architecture is certainly needed, which enables these processes to be handled in an integrated manner.

There are several cognitive architectures concerning the interaction between perceptual/cognitive and motor abilities, however, we apply Model Human Processor with Realtime Constraints (MHP/RT) proposed by Kitajima et al. [7][8][9] for this study. MHP/RT is a cognitive architecture, which is

constructed by extending the concept of Two Minds [10][11] to reproduce the perceptual, cognitive, and motor processes as well as memory processes at work in everyday action selection. MHP/RT has been applied to the comprehension of language utilization [12] and the process of creating ceramic artworks [13]. For the latter study, MHP/RT is applied with a companion field study methodology called Cognitive Chrono-Ethnography (CCE) [9][14]. CCE is a research methodology utilized to clarify the process of development concerning how a specific individual has acquired the behavior selection characteristics at the present time, and the development process of the behavior selection characteristics at the site where the behavior is executed based on the behavior selection mechanism on a time axis, which is specified by MHP/RT. The implementation of CCE requires appropriate research participants – elite monitors – who are ideal for the purpose of the particular research.

In this article, we attempt to understand the process of proficiency in music performance by applying CCE, underpinned by the MHP/RT's underlying concept of Two Minds, such as the interplay between the unconscious process of System 1 and conscious process of System 2. In Section II, starting from an outline of MHP/RT and its fundamentals, MHP/RT's basic processes constituting proficient performance are described. In Section III, the cognitive process in piano performance proficiency based on MHP/RT is modeled, which provides the basis of CCE. In Section IV, the process of proficiency in music performance for selected elite monitors is described. In Section V, the cognitive process model and actual behavior in performance proficiency is compared, and the points that can be interpreted by the model, the points that cannot be interpreted by the model, and the implications from the MHP/RT perspective are thoroughly discussed.

II. Interplay between System 1 and System 2 in Performance Development

The purpose of this study is to understand the stages of practice that a proficient piano player must follow when attempting to complete a piece. The basis for this understanding is provided by a cognitive architecture that allows for an integrated treatment of the perceptual, cognitive, and motor (PCM) processes that take place during practice, as well as the memory processes involved in knowledge use by the PCM processes and knowledge acquisition as the result of the PCM processes. In this section, we first look at MHP/RT, which is the cognitive architecture employed in this study, in Sections II-A and II-B. It then describes in Section II-C the characteristic patterns of execution of PCM and memory processes exhibited by proficient performers, as revealed by previous research [13], which provide a basis for understanding the practice stages of proficient piano players discussed in Sections III and IV.

A. MHP/RT as an Extension of Two Minds

1) Two Minds: Cognitive processes map perceptual information to motor information. Cognitive processes include intuitive and sensory unconscious feedforward control processes that directly map perceptual information to motor information. In addition, there is a conscious feedback control process where the information necessary for mapping, such as information directly connected with perceptual information, secondary information connected with that information, and so on, and finally connected with motor information, is sequentially and serially extracted from memory.

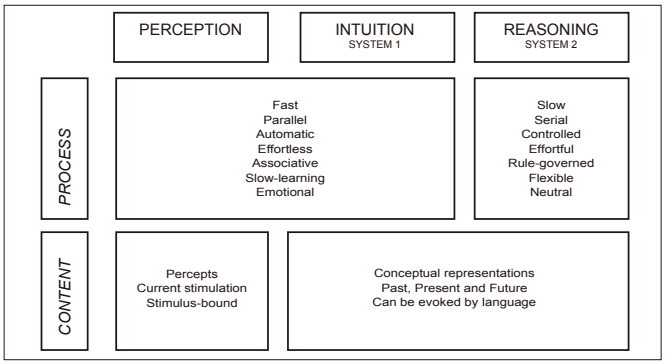


Figure 1. Two Minds [10, Figure 1].

Figure 1 shows the process of situational judgment and subsequent response selection (decision making) in more detail. This figure shows Two Minds proposed by Daniel Kahneman, who won the Nobel Prize in Economics in 2002. Two Minds is the idea that human decision making is carried out by two cognitive systems: System 1, which controls intuition, and System 2, which controls reasoning [11]. System 1 is a fast feedforward control process driven by the cerebellum and oriented toward immediate action. Experiential processing is experienced passively, outside of conscious awareness; one is seized by one's emotions. In contrast, System 2 is a slow feedback control process driven by the cerebrum and oriented toward future action. It is experienced actively and consciously; one intentionally follows the rules of inductive and deductive reasoning.

2) MHP/RT as Two Minds in the Real Dynamic Environment: Two Minds shows that the decision-making process involves conscious and unconscious processes, but it does not say much about the dynamic processes leading up to the decision or the cognitive activities involved in the outcome of the activities performed as a result of the decision. Also, although memory is involved in what is done by the Two Minds process, it says nothing about how memory is used for decision making and updated while reflecting on the outcome of decisions. Daily life can be viewed as a series of behavioral choices involving decision-making, but to understand it on the basis of Two Minds, we need a framework that can provide answers to the unspoken points mentioned above.

MHP/RT directly addresses these points by introducing the idea described below. Considering the human action selection process as a series of perceptual, cognitive, and motor subprocesses, one part of action selection with the processing flow of "perception $\Rightarrow$ cognition by System 1 $\Rightarrow$ action"

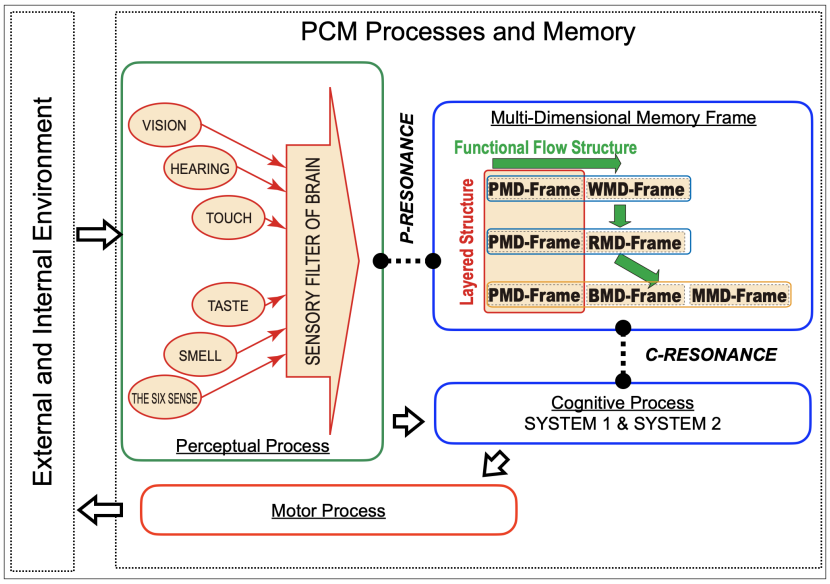


Figure 2. Information uptake by perceptual processes from the external and internal environment, memory activation and execution of cognitive and motor processes through resonance [15, Figure 1].

can be characterized as a *feedforward* control process, in which the unconscious association of perception and motion by intuition leads to action; another part with the processing flow of "perception ⇒ cognition by System 2 ⇒ action" is characterized as a *feedback* control process.

Feedforward control refers to doing things with momentum. In other words, it is a state in which things are done one after another without evaluating what has been done at each point in time (whether it went well or not as expected). On the other hand, feedback control is a state in which the behavior is evaluated each time, the deviation from what was expected is evaluated, and the behavior that reduces the degree of deviation and achieves the expected state is selected from among the actions that can be performed at that point in time, and the behavior is advanced.

Since feedforward control does not involve a cognitive process to evaluate the results of execution, it can link perception and action several times faster than feedback control. Feedback control requires cognitive processes such as understanding the state, evaluating deviations from expectations, and selecting actions that contribute to reducing deviations, e.g., application of the hill-climbing heuristic for solving a problem. Let N be the number of steps required to execute these cognitive processes, the time required is approximately $70 \times N$ msec, where 70 msec is the cycle time required for perming a simple cognitive task, e.g., comparison of two digits. If $N = 10$, it takes 700 msec. Assuming that perceptual unit cycle time is 100 msec and motion unit cycle time is 70 msec, in the case of feedforward control, perception, cognition, and motion can be executed in as little as $100 + 70 + 70 = 240$ msec, e.g., press a button when a circle appears on the display. On the other hand, if feedback control is included, the time required is $100 + 700 + 70 = 870$ msec, which is about three times longer, e.g., press a button when a next move is selected while solving a puzzle after mentally examining a number of alternative moves. Note that in these rough evaluations, the values for perceptual, cognitive, and motor unit times were those reported in the literature as the respective cycle times of the Model Human Processor [16].

MHP/RT is a real dynamic brain model with Two Minds at its core. System 1's unconscious processes with feedforward control and System 2's conscious processes with feedback control are autonomous systems and work together. *both* System 1 and System 2 receive input from the perceptual information processing system in one way, and from the memory system in another way. The cognitive system of Two Minds, System 1 and System 2, are connected with the perceptual and motor systems, and the memory system. These systems work autonomously without any superordinate-subordinate hierarchical relationships but interact with each other when necessary.

B. *Fundamentals of MHP/RT*

The key for understanding the human–environment interaction based on MHP/RT is the idea that the communication between the autonomous systems is achieved by a mechanism of *resonance* [17]. Both environmental systems and human systems are autonomous systems. The human systems as modeled by MHP/RT include the perceptual, cognitive (Two Minds), motor, and memory systems, all of which are autonomous systems. This section describes how the processing of System 1 and System 2 behind continuous the human–environment interaction is supported by resonance mechanisms that link between autonomous systems.

1) Interaction with the Environment Through Memory, Perception, Cognition, and Motor Processes Using Resonance: When interacting with the environment, humans respond to

physical and chemical stimuli emitted from the external and internal environment by sensory nerves located at the interface with the environment and take in environmental information in the body. The brain acquires environmental information concerning the current activity of the self through the multiple sensory organs. Further, it generates bodily movements that are suitable for the current environment. The stable and sustainable relationship between the environment and the self is established through continuous coordination between the activity of the self and the resultant changes in the environment, which should affect the self's next action.

Figure 2 shows the process of MHP/RT [7][9], by which environmental information is taken into the body via sensory nerves as M-dimensional information, processed in the brain, and then acted upon by the external world via motor nerves as N-dimensional movement [15]. This process involves memory, which is modeled as Multi-Dimensional Memory Frame, and perceptual, cognitive (Two Minds), and motor processes. The cognitive process essentially converts the M-dimensional sensory input to the N-dimensional motor output, which is called $M \otimes N$ mapping, with the help of memory. The memory structure, Multi-Dimensional Memory Frame (MDMF), consists of Perceptual-, Behavior-, Motor-, Relation-, and Word-Multi-Dimensional Memory Frame (abbreviated hereafter as P-MDMF, B-MDMF, M-MDMF, R-MDMF, and W-MDMF, respectively). P-MDMF overlaps with B-, R-, and W-MDMF, for spreading activation from P-MDMF to M-MDMF in an attempt to establish $M \otimes N$ mappings.

Perceptual information taken in from the environment through sensory organs resonates with information in the memory network structured as MDMF, which is called *P-Resonance*. In Figure 2, this process is indicated by •—•. Resonance occurs first in the P-MDMF to activate the memory networks. After that, the activity propagates to the memory networks that overlap the P-MDMF, which are the B-, R-, and W-MDMF, and finally to the M-MDMF. In cognitive processing by Two Minds, conscious processing by System 2, which utilizes the W- and R-MDMF via C-Resonance (the upper and middle layers of MDMF), and unconscious processing by System 1, utilizing the B- and M-MDMF via C-Resonance (the bottom layer of MDMF), proceed in an interrelated manner. The motor sequences are expressed according to the M-MDMF, which is the result of cognitive processing. The memories involved in the production of a behavior are updated to reflect the traces of its use process and influence the future behavior selection process.

2) Four Operation Modes: In MHP/RT, the action selection process is controlled by System 1 and System 2 of Two Minds [11]. These systems cooperate to link perception and movement, and the degree of cooperation depends on the state of the external environment with which the MHP/RT interacts. Table I shows the Four Operation Modes characterized by the relationship between System 1 and System 2. There are synchronous and asynchronous modes. Since the activities addressed in this study are concentrated activities, they are performed primarily in the synchronous modes, which are System 1 driven mode (Mode 1) and System 2 driven mode (Mode 2).

TABLE I. FOUR OPERATION MODES OF MHP/RT AND THEIR RELATIONSHIP WITH THE FOUR BANDS IN THE TIME SCALE OF NEWELL'S HUMAN ACTION [18, FIGURE 3-3]; B-, C-, R-, S-BAND REFERS TO BIOLOGICAL, COGNITIVE, RATIONAL, AND SOCIAL-BAND, RESPECTIVELY, ASSOCIATED WITH THE CHARACTERISTIC TIMES, RAGING FROM 10^{-4} TO 10^{7} SECONDS.

Synchronous Modes
Mode 1: System 1 driven mode *A single set of perceptual stimuli initiate feedforward processes at the B- and C-bands to act with occasional feedback from an upper band, i.e., C-, R-, or S-bands.*
Mode 2: System 2 driven mode *A single set of perceptual stimuli initiate a feedback process at the C-band, and upon completion of the conscious action selection, the unconscious automatic feedforward process is activated at the B- and C-bands for action.*
Asynchronous Modes
Mode 3: In-phase autonomous activity mode *A set of perceptual stimuli initiate feedforward processes at the B- and C-bands with one and another intertwined occasional feedback processes from an upper band, i.e., C-, R-, or S-bands.*
Mode 4: Heterophasic autonomous activity mode *Multiple threads of perceptual stimuli initiate respective feedforward processes at the B- and C-bands, some with no feedback and others with feedback from the upper bands, i.e., C-, R-, or S-bands.*

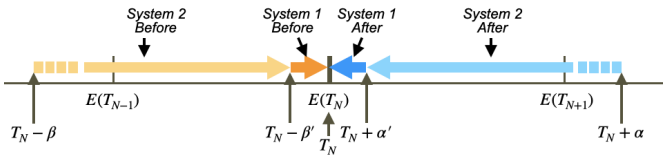


Figure 3. Four processing modes of MHP/RT [13, Figure 3].

3) Four Processing Modes: The experience associated with an individual's activity is characterized by a series of events that are consciously recognized serially. Let $E(T_N)$ denote the event that occurred at time T_N. The experience is then defined as a series of events along the timeline as follows:

$$\cdots \rightarrow E(T_{N-1}) \rightarrow E(T_N) \rightarrow E(T_{N+1}) \rightarrow \cdots$$

Considering the way System 1 and System 2 are involved in individual events, four processing modes can be defined as shown in Figure 3.

- **System-2-Before-Event-Mode**: In the time range of $T_N - \beta \leq t < T_N - \beta'$, MHP/RT plans for future events to occur. There is enough time to think carefully.
- **System-1-Before-Event-Mode**: In $T_N - \beta' \leq t < T_N$, the action selections smoothly generate the immediate event.

- **System-1-After-Event-Mode**: In $T_N < t \leq T_N + \alpha'$, to perform better for the same event in the future, the connection between the incoming perceptual information and the output motor content is adjusted unconsciously.
- **System-2-After-Event-Mode**: In $T_N + \alpha' < t \leq T_N + \alpha$, the event is reflected upon. The results are stored and used in the next System-2-Before-Event-Mode before a similar event occurs.

C. Basic Processes Constituting Proficient Performance

Proficient performance can be understood in terms of the combinations of the four operation modes and four processing modes of MHP/RT when attempting to find $M \otimes N$ mappings by utilizing MDMF. In the previous study [13], we identified the elemental processes that constitute the ceramic artist's skilled work by conducting a CCE study. CCE is a method for obtaining an ecological understanding of how action selection is performed in the domain under study. CCE identifies several parameters that characterize action selection by building a model that can simulate action selection at a coarse level on a cognitive mechanism. Experimental collaborators corresponding to the characteristic parameter value combinations are then selected as elite monitors to observe the action selection process and revise the model based on the results. The perceptual, cognitive, and motor processes, and the memory acquisition and utilization processes that characterize the processes of skilled performance are expected to be common to the piano performance activities discussed in this study. Therefore, in this section, we present the three elemental processes identified in the previous study [13] in a generalized, domain-independent form. They will be related to the specific examples described in the subsequent sections, Sections III and IV.

1) Master Planning in Mode 2: Skilled ceramic work included essentially a conscious decision-making process carried out by the System 2 driven mode (Mode 2 in Table I) for accomplishing the purpose of the ceramic steps such as forming the rough image of the work, selecting the material, and selecting the size. It starts with a respective initial idea followed by an evaluation-update cycle of the idea. It terminates when an idea is evaluated satisfactory.

(β —— β' $-*-$ α' —— α) *Repeat*
- β: Consciously clarify the policy for updating the current idea.
- β': Spread activation in the MDMF.
- $*$: Decide on an update for the current idea.
- α': Organize activation in the MDMF.
- α: Consciously evaluate the updated idea.

Figure 4. Master Planning in Mode 2.

Any skilled activity includes its own set of essential conscious-decision steps for the work. Figure 4 schematically illustrates what is happening in each step in terms of the characteristic moments of the four processing modes, i.e., β, β', $*$, α', and α. At β, a conscious activity starts for the future event to be carried out at $*$ as a consciously recognizable event, i.e., a decision is made. At α, the event is consciously reflected. During the period of (β', α') of several hundred milliseconds, unconscious activities related with the event are carried out.

Each step starts at β for performing conscious reasoning to elaborate the current idea, which could be the initial idea for the step or the updated idea of the previous evaluate-update cycle. The spreading activation within the MDMF proceeds through a series of divergences starting at β' to extend unconsciously the possible paths for establishing $M \otimes N$ mappings, followed by the moment of decision on the updated idea at $*$, where the possible paths are narrowed down to those oriented to a same direction, and the period for convergences terminating at α' to organize them unconsciously along the decided one. Afterward, the decision is evaluated at α. This process is repeated until a satisfactory evaluation for the current idea is obtained. The result constitutes part of "master plan of the work." The decision concerning the master plan is consciously retrievable in the subsequent stages.

The content of the master plan of the work is affected by the extent to which activity is propagated within the MDMF during the period leading up to it. This process is characterized by the richness of the MDMF, or the amount of experience concerning the work. It is carried out by initially placing a seed that represents the image of initial idea *consciously* in the P-MDMF. It ultimately leads to the event concerning a final decision, which is a conscious representation in the W-MDMF referring to a rough image of the work represented in the P-MDMF. This is done by spreading activation in the MDMF, which has been constructed through extensive $M \otimes N$ mapping experience. The final decision for master plan, which is consciously accessible in the subsequent stages, is obtained as activated patterns of the network in the MDMF centered on the P-MDMF.

2) Two System 1-Driven Activities: There are two types of activities conducted under System 1 driven mode, which are conducted to implement the master plan specified in Section II-C1 represented as symbols in W-MDMF. The first one is characterized by extensive unconscious exploration of candidate $M \otimes N$ mappings at a detailed grain size to find the best one (see Section II-C2a), and the second is characterized by initiation of a long unconscious and non-interruptable activities in the real world, i.e., execution of a sequence of unconscious bodily movement represented in M-MDMF (see Section II-C2b).

a) Testing Ideas in Mode 1: In the skilled ceramic work, there is a step to create the modeling manually that will not break in the next firing process, which is an irreversible and uninterruptible firing process and finalizes the modeling as something permanent in the real world described in Section II-C2b. In any skilled activity, there are steps for concretely imagine body movements to implement the master plan. Figure 5 schematically illustrates the characteristic of these steps. It is a repetitive procedure, which is the same as

($\beta - \beta'$ ——//——$*$——//—— $\alpha' - \alpha$) *Repeat*

β: Consciously think of the final performance.
β': Spread activation in the MDMF.
$*$: Select a candidate bodily movement for the final performance.
α': Conduct a trial for the candidate bodily movement.
α: Test if the result of trial matches the final performance.

Figure 5. Testing ideas in Mode 1.

Figure 4; it is different in terms of the longer unconscious period carried out by System 1 before and after $*$.

At β, *perceptual* representation of a candidate for the complete form of performance is placed consciously in the P-MDMF as a seed. Then, $M \otimes N$ mapping is carried out during the period of (β', $*$) to obtain a candidate motor movement of body parts in the M-MDMF at $*$ for producing the performance defined in the master plan; The event that occurred at $*$ is the event that a candidate for a consciously accessible motor action has been selected in the future. During the period of ($*$, α'), the movement represented in the M-MDMF is used to generate *trial* movement. Its outcome is a perceptual representation in the P-MDMF, which is used to make a judgment at α whether it matches the final performance defined in the master plan of the work. If it fails, the outcome might be used as a next seed in the P-MDMF to spread activation in the MDMF. This updating process is repeated until a satisfactory one is obtained.

b) Embodiment of a Series of Actions in Mode 1: There is a step for embodying the trial movement obtained in the procedure described in Section II-C2a. Figure 6 schematically illustrates the procedure for this step. This is a one-time procedure; once it is initiated, it proceeds until it ends.

$\beta - \beta' - * - \alpha'$ ————//———— α

β: Clarify the final performance and a candidate is placed in the P-MDMF.
β': Spread activation in the MDMF.
$*$: Make a decision with the performance and carry it out.
α': Organize activation in the MDMF.
α: Evaluate the decision consciously.

Figure 6. Embodiment of a series of actions in Mode 1.

At β, the final performance is consciously clarified in the MDMF specified in the master plan. Then, the perceptual representation for the trial performance for the master plan obtained in the preceding step is placed in the P-MDMF, followed by $M \otimes N$ mapping from there into the MDMF to have the M-MDMF get activated, which specifies candidate movements of body parts for performance. Unconscious $M \otimes N$ mapping is carried out during the period of (β', $*$) for making decisions on the single sequence of bodily movements

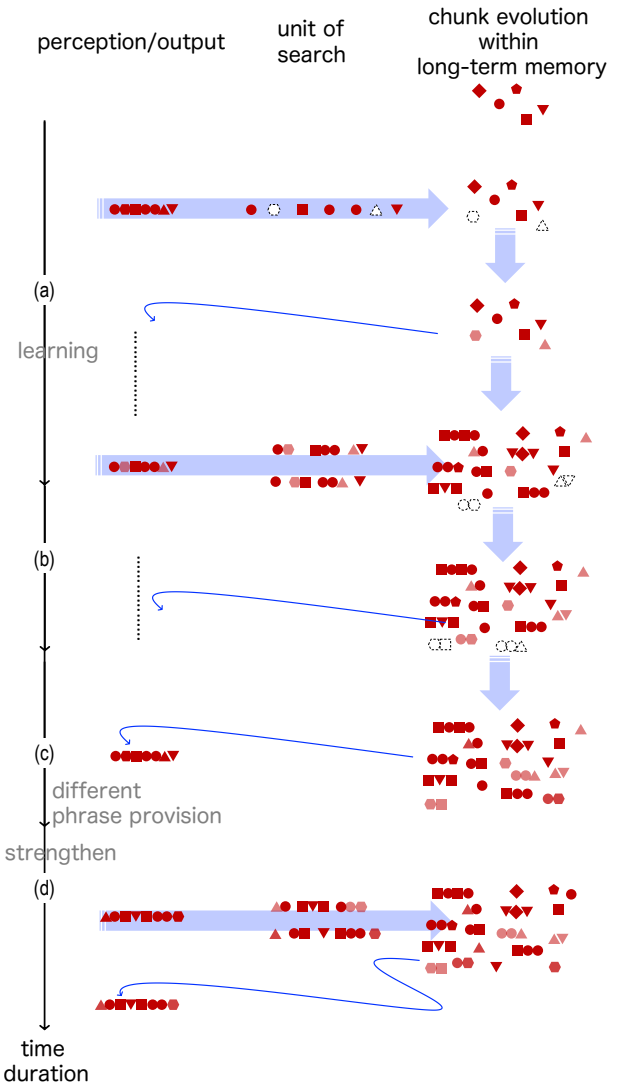


Figure 7. The relation between perception/output and chunk evolution within long-term memory.

in the real world for performance at $*$. This is the moment when the chain of body movements for the performance is established in M-MDMF and the performance is ready to carry out under unconscious feed-forward control by System 1 processing.

After a certain amount of period shown by —//— , the result of the performance will be evaluated at α by System 2. There might exist discrepancies between the final performance imagined at β and the resultant performance obtained at α. By integrating the traces of spreading activation from β to $*$ for performance and the evaluation result at α, the MDMF, which can be used in the $M \otimes N$ mapping for the future embodiment step, is updated.

III. Cognitive Processes Leading to Proficiency in Piano Performance

Playing piano involves processes such as reading the score and creating its mental representations and retrieving knowl-

edge from long-term memory related to the representation, which comprise a variety of information necessary to establish links between the representation of visual information on the score and the concrete hand/finger movements to be conducted on the instrument. These links are used to carry out the $M \otimes N$ mappings introduced in Section II-B1. Long-term memory consists of chunks for establishing these links, which develop with practice from an initial configuration with inefficient linkage to an advanced one with effective linkage, corresponding to the state of proficiency. This section provides a theoretical description for the development process of the chunk structure.

A. Initial State: Initial Chunks in Long-Term Memory

The chunk structure, within long-term memory at the beginning of reading a score, is a set of chunks that have been acquired as knowledge and stored in long-term memory. Let C_{mus} be the chunk set that must be stored, the chunk set C_{LM} that exists in long-term memory at a certain time t is a subset of C_{mus}. C_{mus} is composed of the following, based on the smallest element c_i $(1 \leq i \leq n_c(t))$:

- Chunks composed of the minimum element $n_c(t)$ only,
- Larger chunks composed of $n_e(t)$ $(1 < n_e \leq n_c)$ minimum elements, without duplication, and
- Still larger chunks consisting of $n_{e'}(t)$ $(1 < n'_e \leq n_c)$ minimum elements, with duplication allowed.

In addition, C_{LM} consists of the relation:

$$C_{LM} \quad = \quad \{\boldsymbol{c} \mid c_i \in C_{mus},\ 1 \leq i \leq n_{LM}(t)\}$$

The internal structure of C_{LM} evolves as a learning and strengthening process as the number of chunks it contains increases with practice.

B. State (a): Recognition of Individual Notes or Short Phrases

When reading a new score of music, the perceived sequence of notes is divided into known notes or short phrases. When the learner encounters an unknown phrase, it is stored as a new chunk. The layer (a) in Figure 7 exhibits this state. A sequence of notes $S(t)$ perceived at t consists of n_p elements. When $S(t)$ is initially read, $S(t)$ is separated by n_p individual chunks c_j, and the score reading process commences. When an unknown element $c_{j'}$ appears, $c_{j'}$ is newly stored in long-term memory (black dashed line in the figure). As the score reading proceeds in this manner, the reading of each n_p element proceeds smoothly, and the newly stored $c_{j'}$ is additionally stored and fixed in memory. This is the timing that determines the size of the dimension that processes the perceptual information, referred to as M in II-B1. In this state, the learner plays these phrases with a pause – each c_j plays with intermittent, so that it can only be played with an awareness of partial cohesion. In other words, the memory network to link the input perceptual information to a series of N dimensional motor movements, i.e., the $M \otimes N$ mapping, has not been established yet.

C. State (b) and (c): Recognizing Multiple Chunks Simultaneously

When a sequence of notes can be recognized as individual notes or short phrases, the same $S(t)$ is perceived, but several c_j are lumped together and recognized as a novel chunk (phrase) in order to play the music significantly smoother. The layer (b) in Figure 7 exhibits this state. When the learner perceives this unknown combination of c_j's as a set, it is stored as a new chunk (black dashed line in the figure). At this time, the size of the chunk is larger than that of the state (a), enabling the learner to perform with an awareness of longer chunks. In order to be aware of the large phrases, training is also conducted to recognize $S(t)$ more reliably by separating the elements of $S(t)$, and c_j's, in various ways. When the learner perceives an unknown c_j combination, the combination is newly stored in the long-term memory (black dashed figure in Figure 7). Through repeated training, the number of chunks (phrases) formed by the combination of c_j that existed prior to the training increases in long-term memory, and the learner's chunk set structure incrementally approaches C_{mus}. Finally, the learner's chunk set structure in long-term memory is reached at the state (c), and the presented sequence of notes can be recognized as a single chunk. If the learner's condition reaches the state (c), the learner's skill is regarded as "acquiring the ability to perform $S(t)$ with proficiency." In other words, the construction of the memory network for $M \otimes N$ mapping at a basic level has been completed at this state. It could be augmented further in the next state.

D. State (d): Efforts toward more Reliable Chunking

When the structure of C_{LM} is saturated, even if a sequence of notes is novel to the user, it can be perceived as a known sequence of notes by devising alternative segmentations for c_j, which is equivalent to activating corresponding paths in the $M \otimes N$ mappings. Assuming that a new sequence of notes $S(t')$ consisting only of chunk groups in C_{LM} is perceived, in this regard, the recognition of $S(t')$ is divided by utilizing the chunk elements in long-term memory. Since all the chunks are known, reading will commence without much effort being required. The layer (d) in Figure 7 exhibits this state. In this case, the chunks in long-term memory are simply strengthened.

E. Summary

As the above state is repeated, more C_{LM} is accumulated in long-term memory, and even when it is presented with a complex piece of music, the user can be confident that "this musical piece can be performed". Therefore, as C_{LM} increases in the fusion described above, the more musical pieces the learner practices, the more proficient the learner becomes, and the more musical pieces the learner is able to perform. However, in actual performance, there are two types of practice: one is to perform without making mistakes even if it takes a longer time, i.e., a phase of musical score reading, and the other is to perform without stopping to have

the audience experience a smooth performance. The former is carried out by following the process schematically illustrated by Figure 5 in Section II-C2a where the player tries to confirm a satisfactory performance; whereas the latter by Figure 6 in Section II-C2b where the player lets the motor movements develop without interruption. The process of utilizing chunks while carrying out the $M \otimes N$ mappings should be different in these cases. The next section describes an example of how the cognitive processes, leading to performance proficiency described above, appears in actual performance proficiency with referring to the basic processes included in proficient performance as presented in Section II-C.

IV. An Example of Proficiency Process of Music Performance by a Proficient Piano Player

In this section, we describe a CCE study focusing on a single elite monitor, following the study conducted by Kitajima et al. [13] to understand the skill of a traditional craft artist and how the skill is passed down from generation to generation, as well as how the process by which a proficient piano player reaches the expected performance level through practice of a given piece of music. We call the elite monitor, i.e., the proficient amateur piano performer, P^3, and consider the situation where P^3 tries to achieve a high level of performance perfection through practice. The characteristics of the score that P^3 is aiming for, i.e., the target score abbreviated as TS, with reference to P^3's performance skill level is elucidated. Subsequently, the study enumerates the elements included in the practice to be conducted to achieve TS, and elucidate the development of the practice over time and the content of the practice elements associated with it.

Here, the role of P^3 is taken by the first author. The core of the CCE analysis – describing P^3's experience – has operated as stated below. In order to avoid a biased analysis, when P^3 made an ethnographic analysis, P^3 asked the instructor the meaning of musical suggestion or cognitive meaning with regard to playing piano training method given by instructor. For representation of the CCE analysis, P^3 wrote down the experience series and the initial proposed model. Subsequently, the other two authors, who are professionals with the CCE, meticulously investigated the proposed model which P^3 proposed. Finally, the authors adopt the representation which all authors judged to be acceptable.

A. Main Objectives of a Skilled Piano Learner

In general, there are two main objectives when an adult learner attempts to acquire proficiency in musical performance.

1) Internal factor, such as genuinely wishing to become proficient for strong motives, e.g., favorite piece of music, wanting to perform it, and select a piece for a competition, etc.
2) External factor, i.e., a piece assigned for a competition or given for practice

It depends on which objective the learner set, but here we target the "to be made best performance at the competition" in 1). In this instance, P^3 can select a piece of his/her own will, but the target performance achievement is to pass at least the regional qualifying round of the piano competition (with a required score is 70/80 or higher), and preferably the regional finals (with a required score is 80/86 or higher).

B. Flow of Music Proficiency to Reach Competition Stage

Figure 8 represents the general proficiency process of a musical performance. Given that it takes a long time, anywhere from six months to one year, to become proficient in a music performance, the most important process is the selection of the music to be performed. Basically, there are two important perspectives of selection with regards to music and performing: whether or not the piece is appropriate for the player's performance skill level, and whether or not the player prefers the piece. However, in the case of P^3 who can participate in competitions, there is a lot of freedom in music selection, which means the performance skill level is not a constraint. Hence, P^3 asked her instructor for several candidate pieces that would be suitable for her own timbre and expressive characteristics. On top of that, P^3 herself selected the music to be performed through the following process :

- Give the score a once-over,
- Try out playing the initial few pages (where most of the music motifs are available), and confirming whether or not they can play the piece to the end, and
- Listen to a professional performance and determine if you can grasp the image of the piece.

This stage corresponds to "Master Planning" described in Section II-C1.

After the piece for competition is selected, the learner practices playing it to the end so that the framework of the piece can be imagined. This stage corresponds to "Testing Ideas" described in Section II-C2a. Then, the learner makes *Analise* with the outcome of practice. Post-completing the *Analise*, she fixes the image that expresses fluent performance, and additional interpretation as well as the necessary skills for performance expression. This stage corresponds to "Embodiment" described in Section II-C2b. Subsequently, she will go to the competition performance. Details of each process are described in the subsections to follow.

C. Details of the Processes and Mapping on Two Minds

1) Score Selection toward Practice: There are various ways to select a music piece for competition. In a competition which is not given a set piece of music and in which the goal is to perform well in the competition qualifying round and the finals, there are a number of points to consider in the selection of the music piece. In addition to selecting pieces and considering the level of difficulty, there are some other selection points. In the case of P^3, the following procedure was utilized to select pieces at an appropriate level.

1) Ask her instructor to list some candidate pieces:
   There are two reasons for this. One is to avoid selecting pieces of an inappropriate level for the competition. The other is to have an outsider recommend a piece

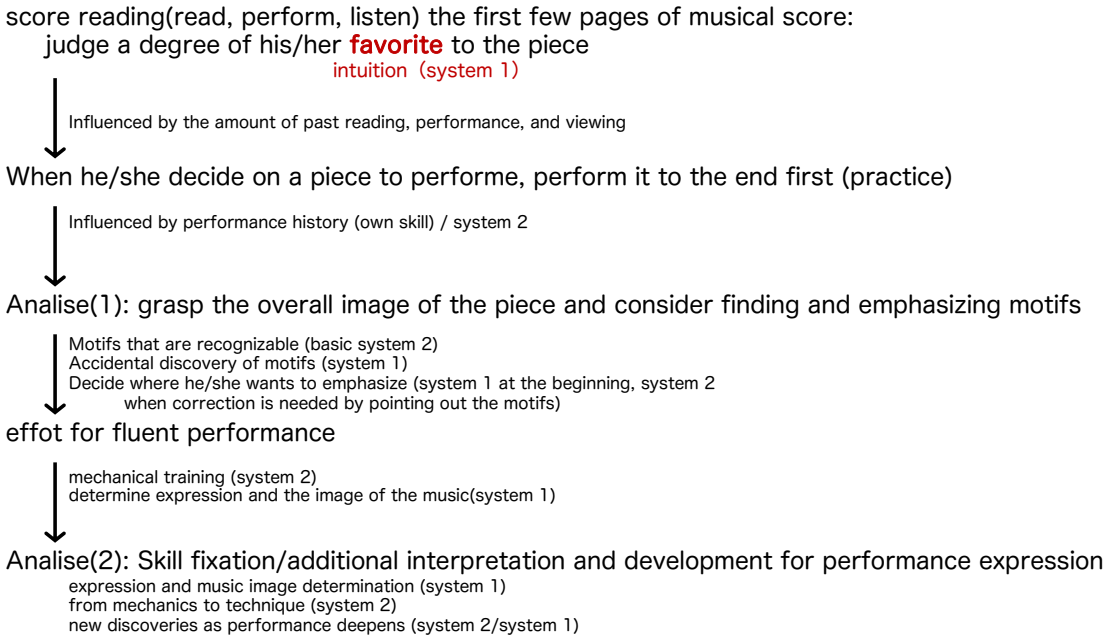


Figure 8. Flow to proficiency in music performance.

that is suitable for the color of P^3 from a third party's perspective.

2) Read the scores giving a once-over to the end to get an image of the music, and narrow down the candidate pieces to 2∼3:
In the case of P^3, the key points in narrowing down the candidate pieces are basically two points: whether the feeling of the music fits, and whether the image of the music can be grasped by reading the scores once-over.
3) Read and perform the initial few pages of the piece (up to the point where the initial and subsequent motifs appear):
There are cases where the mechanics utilized in the actual performance are quite different from the image. In addition, even if the instructor thinks "She can perform this," P^3 finds later that the motifs involves the mechanics "her cognitive or motor reaction rejects." This process is designed to prevent such mismatches.
4) Select a piece of music that she is convinced she could perform well.
In the case of P^3, the selection is made focusing on the music that immediately comes to mind concerning "what she wants to express" when the motifs are performed.

The goals of these steps are 1) listing the candidate pieces, 2) selecting pieces with graspable images, 3) selecting pieces involving mechanics her cognitive and motor reaction accept, and 4) selecting a piece she is confident that she can play. These goals are consciously set and accomplished after several iterations until satisfactory results are obtained as shown in Figure 4. The selected piece for the proficient piano player would correspond to the master plan for the ceramic artist. Both are associated with the images of finished performance via rich memory networks.

2) Transition of Instructional Contents: The process required to complete a musical performance can be divided into two main categories: musical score reading and compositional expression. The musical score reading is a practice stage in which mechanics – motor system – play a major role. This stage is carried out by the process shown by Figure 5. Compositional expression practice is the stage, where musical interpretation, i.e., the player's expression tailored to his/her sensitivity, and technique for the expression, plays a major role. This stage is carried out by the process shown by Figure 6. There are significant disparities between the two practices.

a) Practice 1 – Musical Score Reading: In the case of musical score reading to train mechanics, the main focus is to be able to strike keys accurately. Therefore, the main task of practice is to reproduce the exact note value, pitch, and interval for each note head. In simple words, the primary focus of practice is to count the lengths accurately, to check the details of pitches, and pitches described in the score, and to check accidentals, articulation marks, ornaments, pedal marks, etc. The utilization of knowledge in this process is basically centered on (a) and (b) in Figure 7, and is mainly a System 2 process for consciously evaluating the results of keystroke execution in terms of practicing to play the sequence of notes exactly as described in the score.

b) Practice 2 – Compositional Expression: Conversely, in the case of the musical score reading for compositional expression, a variety of control with regards to the fingers and cognition is required, such as how far to play a note sequence as a whole, how to add dynamics, and which notes to insist on. Simply put, it is a prerequisite that the player has already

finished *Analise* the piece and that the player's image of the entire piece has been established. The two elements are not independent, which means the existence of accurate mechanics enables the player to confidently express music utilizing this technique.

c) Advanced Mechanics Learning for These Practices: It is also necessary to learn the mechanics required to make the technique more precise, for example, the dynamic technique and the techniques required to change timbre. In this sense, it is a cooperative activity between cognitive and motor processes. When teaching these cooperative activities, the instructor decides on the contents of instruction in the following manner with regards to listening to the player's performance.

- Understand what the player wants to emphasize and what kind of expression he/she wants to express from the performance.
- Imagine what the player wants to do but does not seem to be able to do.
- Point out obvious deviations from the interpretation of the performance as described in the score, and give a more natural interpretation.

Of course, if the player is sufficiently competent, these items can be improved in a self-regulating way by recording and watching his or her own performance. However, different from students who are beginners when it comes to performing, there is a limit to self-regulation improvement in the field where *advanced* performance is required. For this reason, suggestions from the instructor play an important role in the case that mastering of advanced mechanics is required for the player who has been already at the level of high proficiency.

The instructor suggests more exercises that would contribute to the formation of chunks as opposed to the movement. They essentially act to strengthen what was weak in the paths connecting perception and movement within the memory network. For instance, changing the playing speed between stressed and unstressed parts (contributes to the formation of chunks), practicing rhythm (contributes to the formation of fingering chunks), and giving more accent than necessary to notes that should be emphasized (contributes to the formation of chunks in the imagery of the music). The primary utilization of knowledge in such exercises is exhibited in Figure 7 (c), and primarily consists of combining the smallest elements c_j that may appear in a piece of music in as long a phrase as possible, in order to be aware of the motifs of the music piece.

d) Improvement through Alternation of These Practices: Given that this is an expression of how the player feels about the music, it is not necessarily a System 2 process, but is gradually shifted to a System 1 process. Repeat the performance expression in the System 1 process as trial and error until the player's intention is well conveyed. The player repeats the pattern that successfully shares the expression he/she wants to share in the System 2 process to fix the expression. In addition, although System 2 and System 1 repeatedly appear during practice, there will be situations where "System 2 $<$ System 1." This is a time when unconscious performances increase and dramatic improvements in performance expressions occur.

As player's technique improves, he/she gradually discovers new discoveries and desires for additional expression in the piece. As player's techniques improve, he/she can make new discoveries for motifs/notes significance, and grow his/her appetite regarding compositional expression. Some of these improvements can be made solely by P^3, while others can only be made with the advice of the instructor. In any case, the final regulation for the competition will be made by repeating such improvements. At this time, the utilization of knowledge increases in the System 1 process in order to challenge a variety of expressions. In addition, even without the System 2 process, the approach to the state known as "the body remembers" and enables various expressions to be challenged.

V. Discussion Based on Two Minds

A. Overview of Annual Lessons

The following is a summary of the practice sessions described in Section IV, contrasted with the duration of the lessons. In order to take lessons, the learner makes practices about one hour per practice. The number of practice sessions is generally two to three times per week, depending on the situation at the time. One to two weeks prior to the competition, practice sessions occurred almost every day.

- 11 months prior to the qualifiers of the competition (C_P): Selection of pieces
  Play a few pages of several music pieces and select the pieces that suit the player's favorite
- Six months prior to C_P post-selection of music pieces: score reading (T_{C1}).
  Basically, the students practice developing techniques in some parts while focusing on the mechanics. It takes about three months to reach the level of playing through the whole piece, and the playing speed is two to four times slower than the specified speed.
- Six to three months prior to C_P: Transition to the expression of musical ideas (T_{C2}).
  ** By this time, the mechanics are 80% complete, so the main focus is on practicing to develop the techniques necessary for compositional expression.
- Three months prior to C_P, completion of the compositional expression:
  Completion of the musical compositional expression · constructing the music image (T_{C3}).
- 1 month prior to $C_P \sim C_P$: final adjustment for the regional qualifying round. (T_{C4}).
- Post C_P to the primary line of the competition: if you pass the qualifying round, practice for the regional finals (T_{C5}).

A total of 25 lessons were given. Each lesson lasted approximately 1.5 hours.

TABLE II. PHASE CLASSIFICATION OF KNOWLEDGE/COGNITIVE PROCESSES AND DEGREE OF INFLUENCE.

		process		knowledge		environment
Phase	Subphase	System 1	System 2	tacit	explicit	outsider intervention
decide piece	offer candidate	*	*		*	**
	once-over	*	**	*	**	**
	playing trial	**	*	**	**	
	listning	***	*	**	*	*
	select piece	***	*	***	*	*
score reading	fingering		***		***	
	score reading		***		***	
analise(1)	recognize motif		***		***	*
	set enphasis		***		***	*
	find motif of serendipity	***	*	*	***	
expression	mechanic	*	***	**	***	***
	construct image	***	**	**	**	
	transfer expression	**	**	*	***	**
analise(2)	confirm expression	**	***	**	***	***
	confirm image	**	***	***	**	***
	Technic		***		***	***
	performance deepening/serendipity	***	***	*	***	
final stage	fragmentation and reintegration	**	**	**	***	

B. Two Minds in the Flow Leading to the Completion of the Music

Once a series of experiences had been performed, the second trial for attending the competition may be able to utilize the prior experience to finish the piece at a faster pace. The items from stage 2 (practice) analise(1) to the effort for fluent performance in Figure 8, or $T_{C1} \sim T_{C2}$ in terms of the lesson schedule, are basically affected by the experience. It is possible to reach the stage of mechanical performance as reproducing with midi, through an experience such as earlier through participating in competitions repeatedly, taking lessons for many years, and so on. These changes are continuous, i.e., the degree of improvement increases monotonically as a function of the number of performances.

However, additional interpretation and deepening of the performance beyond that point may not be successfully achieved by simply repeating the process. In P^3's participation in the competition, the performance around two to one month prior to the competition qualifier (T_{C3}) undergoes a large change every year, which cannot be explained by the passage of time alone. By this time, the mechanical performance is almost complete in a form that is approximately 1.5 times less than the speed at which it is played on the day of the competition, but it is far from sufficient completion, and the so-called "composition expression and understanding." Around the transition from T_{C2} to T_{C3}, there is a significant change in the recognition of musical motifs and a shift to the recognition of larger motifs and the expression of *Dynamik* including expression marks. Other changes in timbre, for instance, from soft to hard sounds, are also observed.

This situation is further analyzed from the perspective of the disparities between the characteristic times of System 2 and System 1. In System 2 driven mode, the processing flow is controlled consciously as shown by Figure 4, whereas in System 1 driven mode it is controlled unconsciously as shown by Figures 5 and 6. In both modes, part of the memory network that connects perception and motion via cognition is used, created if necessary, and the connections are updated, which concerns usage and maintenance of $M \otimes N$ mapping. The period of T_{C1} is a practice process in which the System 2 process is dominant. The time scale for practice per phrase is primarily the cognitive band in Newell's Time Scale of Human Action [19], since the phrase itself is not very long. The time span of the cognitive band is about $\sim$10[s]. Given that information is exchanged between the working memory and long-term memory in about 10 seconds of very short chunks, all knowledge is likely to be recognized only as fragments. Therefore, even if one were to predict the next chunk that will appear during the performance of a piece of music, only a few chunks exist which is able to collation, and even if many chunks can make connected collation, only a few percent of the entire piece can be predicted, making it difficult to see the entire piece.

By repeatedly practicing a very short chunk, the body remembers new chunks in the order of ease with regards to memorizing. If a similar chunk had been utilized in the

past, it is recognized as a "meme" and the chunk becomes an active meme [20]. At this stage, the chunk is considered an action-level meme. Conversely, even if a chunk exists in long-term memory, if it is never invoked again, the chunk is no longer imitated and becomes an extinct meme, therefore making it inactive. From the above, for a learner like P^3 who cannot engage in constant piano practice, score reading at the competition level will require an enormous amount of time.

However, by the time the T_{C1} period had elapsed, the information per chunk is considerably larger. Therefore, during T_{C2}, chunks of the larger size are available for the cognitive processes in the cognitive band. The number of chunks available for cognitive process, invoked chunks, is getting longer and longer, and their coverage is getting longer. As a result, the number of operations utilizing the working memory and long-term memory for a unit time will be gradually increased, and the addition of information to the chunks in long-term memory will be accelerated. In simple words, it is thought that the easily accessible active meme will change to behavior-level meme [20]. In this process, the time when a knowledge group is composed of only an appropriate chunk size may be approximately the time toward T_{C3}.

By the time T_{C3} is entered, the number of movements to call chunks from long-term memory is considered to be considerably reduced. As a result, cognitive-motor coordination is conducted more unconsciously. If all the chunk invocation patterns are optimized, almost all the performances will be performed unconsciously by System 1, and an abrupt phase transition from the T_{C2} state will occur. As a result, one should feel at least a dramatic improvement in their ability for good finger movement.

In the case of P^3, the pieces learned in the last three years, including the time of writing this article, were as follows:

- 2 years ago :
  Partita BWV 826, composed by J. S. Bach (score A)
- 1 years ago :
  Allegro Appassionate op.70, composed by Charles Camille Saint-Saëns (score B), Allemande in French Suites BWV 812, composed by J. S. Bach
- now :
  piano sonata op. 14 first movement, composed by Sergei Sergeyevich Prokofiev (score C), Allegro in Italian concert BWV 971, composed by J. S. Bach

Each of them spent about a year memorizing the scores prior to the competition. Despite the difference in the compositional age, compositional structure, and knowledge required, score A received 76 points and score B received 79 points in the final piano competition. This indicates that the learners' performance skills themselves were well-developed, even though they performed different types of music. In simple words, the examples of the experience in Section IV can be considered to have a certain universality.

C. The Relation between Knowledge/Cognition Process and Two Minds

Finally, we discuss the relationship between the Two Minds and the knowledge as well as cognitive abilities acquired through a series of piano practice and lessons. Table II exhibits the results of subjective evaluation for each flow subphase in Figure 8. The items are: the process of the Two Minds, the knowledge system utilized (implicit/explicit), and the subjective evaluation of the degree of intervention by others. The higher the number of $*$, the stronger the effect on the item.

At initial glance, one might think that instrumental music performance is a continuous shift from long time-consuming mechanical training by System 2 (inference) to unconsciousness of musical expression including System 1 (intuition). However, in fact, this is not true.

For instance, in the case of the music selection phase, many factors are involved in the decision-making process, including player: 1) preference (System 1), 2) matching with performance ability (System 1/2), and 3) matching with the ability to read music (System 2), etc. It depends on the situation at that time which of these factors should be prioritized. In simple terms, if motivation is a given priority, preference is given priority, and if ability is given priority, a little more weight is given to the performance ability or reading ability. This indicates that the process of proficiency in instrumental performance is not determined solely by preference or ability. Conversely, music selection, although often neglected at the initial glance, is the most important phase as it is deeply related to the motivation of the student when he or she begins to practice. In the case of the piano beginner, the instructor often selects pieces at an appropriate level, but in the case of a proficient amateur learner, the selection requirements for the score selection are reduced to some extent. Therefore, the degree of freedom of parameters is high, and the decision-making process involves a mixture of perceptual processes to trigger preference by listening to the sound source, perceptual-cognitive processes to compare with the reading ability by score reading, cognitive-motor processes to consider the performance ability, and processes to coordinate all of these. Therefore, the ability to select appropriate music can be regarded as an important ability.

This also applies to the score selection process. It is easy to assume that a System 2 process takes precedence in *Analise* as well, since it requires a precise analysis of the music. However, various cognitive processes are intricately related as follows: Recognizing the motive and searching for methods to emphasize it (System 2), determination of the expression method that is perceived as effective (System 2/1), new expressions discovered by chance (System 1), and so on. Therefore, not only an orderly musical interpretation but also a balance with the impression is important. In particular, when representing a piece of music, it is necessary to "see the big picture", i.e., the following items must be fulfilled at the same time.

- The player must have a complete understanding of how to perceive the entire piece (System 2).

- The player's natural behavior as if he/she were performing it unconsciously, which should be comfortable for the listener (System 1).

Therefore, it is necessary to understand the process of coordination between System 2 and System 1.

VI. Conclusion and Future Works

In this study, based on the MHP/RT cognitive architecture and its companion field study methodology, CCE, we attempted to understand the process of proficiency in music performance by proficient piano players as a brain model based on the coordination of perception, cognition, and movement, as well as the Two Minds.

In Section II, we reviewed MHP/RT and provided its fundamentals that are needed to understand the process of training and performance of proficient piano player in preparation for competition. It identified three basic processes constituting proficient performance, which served as the elements for the description hereafter.

In Section III, we theoretically explained the development process of the chunk structure that exists in the long-term memory, which is the most important part of the piano playing process – score reading and piano playing mechanics/technics. There is a structure, which consists of many small units of chunks in the long-term memory, and links are attached between chunks through practice. As a result, larger chunks are formed. The study argues that the proficient state refers to this state.

In Section IV, we ethnographically described the piano practice and proficiency process with P^3 as an example, aiming at participation in the competition. We exhibited that there are four major components: selecting score (System 1), practice (System 2), *Analise*(System 1/ 2), and the effort for fluent performance (System 1/2).

In Section V, a series of piano playing exercises and lessons were analyzed from the perspectives of the Two Minds, the knowledge system utilized (implicit/explicit), and the intervention of others. Post the analysis, the relationship between the acquired knowledge and cognitive abilities as well as the Two Minds was examined by incorporating the idea of the active meme. The results suggest that instrumental music performance requires both a complete understanding of how the player perceives the entire piece (System 2) and natural behavior that is comfortable for the listener (System 1), as if the player were playing unconsciously.

As an application, we can consider various educational support measures for performance proficiency by understanding the actual growth process of chunks and the player's proficiency process in more detail based on cognitive architecture. In recent years, there have been increasing opportunities for adults who are not professions of instrumental music performance to enjoy music as a hobby as amateurs. While he/she is not a professional with regards to instrumental performance, one of the elements necessary for proficiency, "motivation to practice" and "support for its maintenance", is left solely to the desire of the learner to play this piece, not to the instructor. In this situation, if learners cannot overcome the difficulties they encounter when practicing instrumental music, they may give up the hobby of instrumental music itself. However, if the instructor can appropriately understand the difficulties that the learner cannot overcome, and can demonstrate to the learner how to increase the possibility of overcoming the difficulties, the withdrawal rate of the learner may be reduced. We believe that this study will contribute to the research from this perspective.

The majority of prior research on the process of proficiency in musical performance has focused on the understanding of cognitive mechanisms for individual perceptual, cognitive, and motor processes. Research on the cognitive mechanisms of individual processes is primarily suitable for understanding proficiency or the process of developing literacy, in terms of how beginners can play music. This study's findings can apply to constructing efficient training methods for the novice learner.

However, learner's playing skill shifts slowly with time, so that it is necessary to improve teaching content and methods based on the learner's proficiency. In case the learner's goal level with regards to attending the competition, is not only the improvement of literacy but also the process of proficiency in the "big picture" of a piece of music. In order to establish such a sophisticated instructional method for individual cases, we need a method for analyzing successful/failed cases based on the empirical rules of instruction, and the resulting cognitive model of the learner. In this case, it is necessary to go into the resonance with past performance and appreciation activities, and there are many areas that cannot be elucidated only by the prior cognitive architecture. As one of the solutions to this problem, understanding performance proficiency utilizing a brain model based on the Two Minds is considered to be effective. As a future issue, we believe that further research based on this study will enable, for instance, remote performance instruction of musical pieces at a higher level.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 19K12232, 20H04290, 22K02840, 22K02885, 23K11334. The author would like to thank Editage (www.editage.com) for English language editing.

References

[1] K. Nakahira, M. Kitajima, and M. Toyota, "Practice Stages for a Proficient Piano Player to Complete a Piece: Understanding the Process based on Two Minds," in COGNITIVE 2024: The Sixteenth International Conference on Advanced Cognitive Technologies and Applications, 2024, pp. 21–29.

[2] C. Palmer, "Music performance," Annual Review of Psychology, vol. 48, 02 1997, pp. 115–38.

[3] A. Lehmann and K. Ericsson, "Research on expert performance and deliberate practice: Implications for the education of amateur musicians and music students," Psychomusicology: A Journal of Research in Music Cognition, vol. 16, 04 1997.

[4] M. V. Araújo, "Measuring self-regulated practice behaviours in highly skilled musicians," Psychology of Music, vol. 44, no. 2, 2016, pp. 278–292. [Online]. Available: https://doi.org/10.1177/0305735614567554

[5] R. Chaffin, G. Imreh, A. F. Lemieux, and C. Chen, ""seeing the big picture": Piano practice as expert problem solving," Music Perception, vol. 20, no. 4, 2003, pp. 465–490.

[6] R. Chaffin and L. Topher, "Practicing perfection: How concert soloists prepare for performance," Advances in Cognitive Psychology, vol. 2, 01 2006, pp. 113–130.

[7] M. Kitajima and M. Toyota, "Decision-making and action selection in Two Minds: An analysis based on Model Human Processor with Realtime Constraints (MHP/RT)," Biologically Inspired Cognitive Architectures, vol. 5, 2013, pp. 82–93.

[8] M. Kitajima and M. Toyota, "Simulating navigation behaviour based on the architecture model Model Human Processor with Real-Time Constraints (MHP/RT)," Behaviour & Information Technology, vol. 31, no. 1, 2012, pp. 41–58.

[9] M. Kitajima, Memory and Action Selection in Human-Machine Interaction. Wiley-ISTE, 2016.

[10] D. Kahneman, "A perspective on judgment and choice," American Psychologist, vol. 58, no. 9, 2003, pp. 697–720.

[11] D. Kahneman, Thinking, Fast and Slow. New York, NY: Farrar, Straus and Giroux, 2011.

[12] M. Kitajima et al., "Language and Image in Behavioral Ecology," in COGNITIVE 2022: The Fourteenth International Conference on Advanced Cognitive Technologies and Applications, 2022, pp. 1–10.

[13] M. Kitajima, M. Toyota, and J. Dinet, "Art and Brain with Kazuo Takiguchi - Revealing the Meme Structure from the Process of Creating Traditional Crafts -," in COGNITIVE 2023: The Fifteenth International Conference on Advanced Cognitive Technologies and Applications, 2023, pp. 1–10.

[14] M. Kitajima, "Cognitive Chrono-Ethnography (CCE): A Behavioral Study Methodology Underpinned by the Cognitive Architecture, MHP/RT," in Proceedings of the 41st Annual Conference of the Cognitive Science Society. Cognitive Science Society, 2019, pp. 55–56.

[15] M. Kitajima et al., "Basic Senses and Their Implications for Immersive Virtual Reality Design," in AIVR 2024: The First International Conference on Artificial Intelligence and Immersive Virtual Reality, 2024, pp. 31–38.

[16] S. K. Card, T. P. Moran, and A. Newell, The Psychology of Human-Computer Interaction. Hillsdale, NJ: Lawrence Erlbaum Associates, 1983.

[17] J. Dinet and M. Kitajima, "The Concept of Resonance: From Physics to Cognitive Psychology," in COGNITIVE 2020: The Twelfth International Conference on Advanced Cognitive Technologies and Applications, 2020, pp. 62–67.

[18] A. Newell, Unified Theories of Cognition (The William James Lectures, 1987). Cambridge, MA: Harvard University Press, 1990.

[19] J. V. Monaco, "Classification and authentication of one-dimensional behavioral biometrics," in IEEE International Joint Conference on Biometrics, 2014, pp. 1–8.

[20] M. Kitajima, M. Toyota, and J. Dinet, "How Resonance Works for Development and Propagation of Memes," International Journal on Advances in Systems and Measurements, vol. 14, 2021, pp. 148–161.

Analysis of Accessibility Problems in Medical Devices

Mariana Ribeiro Brandão
Institute of Biomedical Engineering (IEB-UFSC)
Federal University of Santa Catarina
Florianópolis, Brazil
e-mail: marianaribeirobrandao@gmail.com

Renato Garcia Ojeda
Institute of Biomedical Engineering (IEB-UFSC)
Federal University of Santa Catarina
Florianópolis, Brazil
e-mail: renato.garcia.ojeda@ufsc.br

Abstract— **The availability of an affordable medical device is critical in the provision of healthcare to ensure that technology is not a barrier for users. It is essential to understand the accessibility issues present in medical devices to serve the diverse population of patients with varying limitations, abilities and disabilities. With the aim of promoting and discussing the impacts of accessibility in medical devices, this project aims to analyze accessibility problems in medical devices. A rapid review of the literature was prepared and a model for applying usability methods throughout the life cycle of health technologies was proposed to establish strategies to improve accessibility and mitigate risks. This work found a large number of accessibility problems involving different types of medical devices, as well as the lack of accessible technologies in healthcare environments. Different actions to provide a more inclusive and accessible health technology management throughout the life cycle were proposed, such as incorporating user-oriented development, training and development of standard operating procedures.**

Keywords-Accessibility; Medical Devices; Health Technology Management.

I. INTRODUCTION

The availability of an affordable medical device is critical in the provision of healthcare to ensure that technology is not a barrier to users [1][2]. To achieve the benefits for which the medical device was developed, it requires a safe and reliable technology-user interaction, so that errors in use by users do not cause harm, compromising the health of the population [3]. Therefore, a combination of human-centered project development, ergonomics, and accessibility tools, is necessary to ensure a high quality use of technological resources [4].

Considering accessibility aspects in the development of health technologies is essential to ensure inclusion and improve usability. Accessibility is defined in ABNT NBR 17060:2022 as follows: accessibility on mobile devices consists of the scope in which products, systems, services, environments and facilities can be used by people from a population with the widest variety of characteristics and capabilities, to achieve a specific objective in a specific context of use [5]. Incorporating usability into the projects aims to expand the target population, making technologies accessible to more people in different contexts of use [6]. In Brazil, the population with disabilities was estimated at 18.6 million (considering people aged 2 and over). The number corresponds to 8.9% of the population in this age group [7]. In the world, this number is estimated at 1.3 billion, representing 16% of the world's population [8]. According to law Nº. 13.146, of July 6, 2015, which establishes the Brazilian law on the inclusion of people with disabilities, every person with a disability has the right to equal opportunities with other people and will not suffer any type of discrimination. In addition, people with disabilities are being guaranteed comprehensive health care at all levels of complexity, with universal and equal access [9].

However, people with disabilities often do not have the opportunity to receive quality healthcare and sometimes have access to insufficient healthcare [10]. As technologies are increasingly present in healthcare, and are incorporated to assist users in their safer and more reliable use, consideration of accessibility aspects in technological development becomes a fundamental requirement to achieve the usability of a product [3]. Incorporating principles and methodologies considering usability and accessibility must be strategic business objectives, being essential to optimize performance, minimize undesirable consequences for human beings, maximize the well-being of the entire organization and improve relationships with customers [6].

The tool used to evaluate human interaction with a product is usability, and its consideration in health is fundamental and useful for evaluating the user experience [11]. Usability is a metric used to measure how much a product can be used by certain users and achieve specific objectives, when considering parameters such as effectiveness, efficiency and satisfaction in a context of use [12]. For a product or process to have good usability, it is necessary to consider different parameters and measure them with the intended users, such as effectiveness, efficiency, satisfaction, use, learning and accessibility. Accessibility is determined by the ease of access to the products necessary to complete the objective by people with the widest variety of capabilities [6][13]. When considering accessibility, it allows clarity and simplicity in design for people who may temporarily have some limitations or those who have them permanently [13].

The development of a product or service centered on the user's needs and perspective, integrated with their context and tasks, is called User-Centered Development [14]. It consists of an approach to developing usable and useful systems in an interactive way, with an emphasis on users when considering their needs, through the incorporation of ergonomic knowledge and techniques. There is a diversity

of usability methods that aim to support human-centered design, used to increase the usability of a product or system, which can be used in both design and evaluation. Some methods consist of: user observations; questionnaires; critical incident analysis; interviews; think out loud; document-based methods, among others [15]. Accessibility must be included as part of the human-centered project, so that it can expand the population that can use technologies effectively, efficiently and satisfactorily, and consequently, increase usability for all users [6].

Healthcare accessibility is essential in providing medical care to people with disabilities. Due to barriers, individuals with disabilities are less likely to receive routine preventative medical care than people without disabilities. Accessibility is not only required by law, but is also crucial for the inclusion of all people in the use of health technologies [2]. Work involving accessibility in medical equipment reinforces the problems surrounding technologies, as presented in the research conducted by Story et al., which showed harm to people with disabilities when using scales, examination tables and diagnostic imaging equipment [10]. Other equipment and description of accessibility problems will be presented in this article in Section III.

Due to the importance of considering accessibility to ensure the inclusion of all people in the use of medical equipment, this work aims to carry out a rapid review of the literature in search of evidence as well as provide a model for incorporating in Health Technology Management.

The rest of the article is structured as follows. In Section II, we discuss the methodology used in the research. In Section III, the results are elucidated. In Section IV, we discuss the results found. Section V concludes the work with a summary and future research directions.

II. Materials and methods

This work was conducted in two stages. The first phase consists of the rapid literature review and the second phase consists of the proposal for a model that incorporates accessibility tools into the life cycle of a medical device in order to contribute to the safe management of health technologies. To explore accessibility in medical devices and discuss the contribution of Clinical Engineering to making healthcare environments more inclusive, a rapid review was carried out in the literature, which consists of a reliable and systematized methodology to synthesize knowledge. This approach is used when steps in the process of a systematic review are simplified to produce information from the selection of research that is available in the literature, and that is relevant to a study topic [16]. The constant increase in the amount of research carried out in the literature requires the implementation of an approach to evaluate published studies and contribute to decision-making, and thus provide an updated summary of the state of knowledge [17].

The conduct of this rapid review was based on the Methodological Guideline of the Ministry of Health for the preparation of systematic reviews [18], as well as on the PRISMA methodology of the University of Oxford, which consists of a set of evidence-based items that aim to assist in the presentation of research results [19]. The guiding question of the rapid review research proposed for this case study was: **"What is the evidence of accessibility issues in medical devices?"**

To answer this question, the search strategy used was through the definition of keywords to identify publications that respond to this theme. The use of the logical operators "AND" and "OR" helped in the literature search. The search in the databases was executed using the union of keywords: (*"medical device"* OR *"medical equipment"*) AND (*"accessibility"* OR *"disabled people"* OR *"disabled person"* OR *"disability"*) during the time period from January until February, 2024. The search was implemented in the following electronic databases: IEEE *Xplore* and Pubmed, which were used systematically. To determine the choice of articles, inclusion and exclusion criteria were established, which included population parameters of the intended technology, the type of intervention used, the availability of the work, the date of publication and the type of evaluation of the results. After the initial search, the date of publication, the titles and abstracts were read, selecting a total of 12 publications. Table I explains the number of articles found per database using keywords.

Table I. Number of articles found per database.

Database	***"medical device" OR "medical equipment")***	***("accessibility" OR "disabled people" OR "disabled person" OR "disability")***	***("medical device*" OR "medical equipment*") AND ("accessibility" OR "disabled people" OR "disabled person" OR "disability")***
Pubmed	38.722	488.575	722
IEEE Xplore	11.512	20.788	139

The second stage of this work was to propose a model that incorporates accessibility during all activities of the life cycle, hence contributing to the Health Technology Management in pre-market and post-market.

III. Results

The results obtained through a quick literature review highlighted accessibility problems in different types of medical devices, such as examination tables [20]-[22], weight scales [23][24], nebulizers [25], glucometers [26], positive airway pressure device [27], neuromodulation devices [28], mammography [29]. The usability techniques applied to explore and investigate the problems were

mainly: questionnaires, interviews, focus groups and usability testing.

In the studies analyzed, it was found that medical devices are often not accessible to the entire population. Story *et al.* highlighted problems faced by patients with disabilities who have difficulties using different types of medical equipment. The four main equipments with the biggest reported problems were: tables; radiology equipment; rehabilitation and exercise equipment and weight scales. Possible physical damage and incorrect reading of display values were the most recurrent problems, followed by physical positioning and transfer of patients on medical equipment [10].

The absence of accessible medical equipment was presented in some studies, such as the research conducted by Morris et al. in outpatient clinics [20], which converges with Mudrick et al. research that found the absence of adjustable exam tables and accessible weight scales in a large part of offices analyzed [21]. Iezzoni et al. showed that doctors do not use accessible exam tables/chairs for patients, and that many doctors simply ask the weight of patients with mobility limitations [22]. Agaronnik et al. presented in her study that medical diagnostic equipment, such as examination tables, scales and diagnostic imaging equipment are often inaccessible. Even if doctors have accessible equipment (e.g., examination tables), they do not always use them [24].

Accessibility in glucose monitoring system

Technologies used by people with diabetes such as glucose monitor and continuous glucose monitoring systems, have presented accessibility problems in the design of the device that can impact on the erroneous administration of medication. The patients and/or health professionals use the results of the devices to make decisions. Some of the problems highlighted in glucometers are low-contrast displays that are difficult to see for people with low vision [39], test equipment without color contrast [40], absence of speech output, small visual display and high levels of reflection [26].

Study conducted by Akturk, highlights many difficulties healthcare professionals face in initiating diabetes technologies in visually impaired patients with diabetes. This calls for restructuring education and industry support for providers to help them successfully integrate diabetes technologies to improve outcomes among challenging patients with diabetes [41].

Recommendations such as having a sufficiently large display and good display quality (good contrast and anti-reflective screen), support for voice handling of the device, tactile markings, and acoustically well-audible output of the measurement result, warnings and alarms should be considered when developing accessible blood glucose monitoring systems [42].

Accessibility in pulse oximeters

Accessibility in medical devices must be considered for all people. Pulse oximeters are a technology that many studies show can overestimate the true oxygen concentrations of these patients, especially at lower oxygen saturations. The overestimation of oxygen saturation in patients has serious clinical implications, as these people may receive insufficient medical care when pulse oximeter measurements suggest that their oxygen saturation is higher than the true value, which can lead to increased mortality [44].

Different retrospective clinical reviews using electronic health record datasets have shown lower accuracy during the use of oximetry and increased bias in patients with dark skin tones, as well as Asians and Indians compared to white patients, increasing the racial and ethnic disparity in health care [43]-[45].

A study conducted by Gottlieb, showed that Asian, Black, and Hispanic patients had higher average SpO2 readings than White patients for a given blood hemoglobin oxygen saturation. They also received less supplemental oxygen when adjusting for potential confounders, and these disparities appear to be mediated by larger discrepancies between SpO2 and blood hemoglobin oxygen saturation [44]. Another study of premature neonates found a racial disparity in the measurement of oxygen saturation by pulse oximetry and an increased incidence of occult hypoxemia in black premature babies [46]. Possible causes may include factors inherent in pulse oximeter design, insufficient calibration of devices in black individuals and inadequate standards for device approval [43].

Accessibility in scale weight

The lack of accessible scale weight available in healthcare facilities that could accommodate a wheelchair or other assistive technology is a reality in healthcare settings [10][22][47]. Where wheelchair scales were not available in the doctor's office or clinic, healthcare professionals often asked patients to estimate their own weight, potentially leading to health implications for the patient [10].

The most common accessibility problems identified for scales involved the positioning of the patient, the location and legibility of the visual display and the capacity of the scale. For people with low vision, the lack of color differentiation and no strong contrast was reported by patients as a safety issue. Visually impaired people were often unable to read the scale's display, so recommendations such as a display with large letters (and high contrast) or audible and Braille output [10].

Accessibility in Dental Chair

There are several barriers faced by people with disabilities during care in dental services [48]. One of them

concerns wheelchair users when using the dental chair, especially when transferring for procedures, are unable to transfer independently. Patients with physical disabilities may have difficulty getting up and down in the dental chair, positioning themselves or keeping their balance during the procedure. Transferring patients from wheelchairs not only requires manpower, but can also create unnecessary anxiety and even accidents [49].

New technologies are being developed so that patients can be treated while remaining in their wheelchairs without any transfer. However, the number of services that have these accessible technologies is still small, and there are other types of disabilities besides those that people with disabilities suffer from [49]. Patients with intellectual disabilities during dental care may have difficulties with communication and comprehension, which makes it difficult to understand instructions and consent to treatment. Cases of people with Alzheimer's who would not like to be transferred and obese patients who may be too heavy for the dental chair are other common accessibility problems [50][51].

In a study conducted by Isaque et al., it was found that of 400 people with disabilities participating in the survey. In this total, 31.5% considered the inability to sit in the dental chair as one of the main access barriers in dental services [52]. In convergence with another study, conducted by Kanvani et al., it addresses the transfer to the dental chair and remaining immobile for a long time as one of the main challenges in addition to other aspects of the environment's infrastructure. Excessive height of the dental chair, intrusive position of the dental chair arm, lack of support devices, material of the chair covering, lack of skills of the dental team in the transfer and positioning process are some other reports related to the problems faced by people with disabilities in the use of dental chairs [53].

With regard to the technologies used in dental care for patients with disabilities, in some cases chemical and/or physical restraints are used to ensure compliance and immobility [51].

Accessibility in Diagnostic Medical equipment

Many difficulties reported by disabled patients with imaging equipment are related to the platform associated with the equipment, such as contact surfaces, transfer support and positioning support [10].

Specifically in relation to Magnetic Resonance Imaging (MRI), frequent reports relate to not being able to take their wheelchairs or scooters into the MRI room (due to the magnetic field) and therefore having to complete multiple, and sometimes difficult, transfers to reach the machine platform [10]. Performing MRI scans on obese patients can also be interfered with due to the capacity supported by the equipment and safely fitting the patient inside the bore [54]. The capacity of the imaging equipment was also reported in the study conducted by Story et al., who reported that patients were unable to have an MRI scan at their health unit because the diameter of the machine was too small to accommodate them [10].

Accessibility in women health technologies

Women with disabilities encounter a number of barriers to receiving clinical preventive services [53][60], are less likely to have a pap smear, mammogram or breast exam [56] as well as face physical access barriers in the detection and treatment of breast cancer and cervical cancer [57][58].

A study conducted by Story et al., reported that women with disabilities had difficulty maintaining positions during gynecological examinations and that some examination tables and auxiliary components did not offer sufficient support to be able to maintain appropriate positions for examinations or procedures [10]. Lack of accessible diagnostic equipment, such as height-adjustable examination tables and mammography machines, problems accommodating and positioning patients, lack of efficient mammography procedures that meet the needs of women with intellectual disabilities with physical and/or psychiatric limitations are among the challenges that impact on the health of women with disabilities [56][58].

The mammography machine was also the target equipment for research. Yankaskas et al., investigated women with visual, hearing, physical or multiple disabilities on reasons for not returning for regular mammograms. She found that women with multiple limitations were much more likely to report problems with transportation, parking, and accessibility to health services, as well as a lack of medical recommendation for screening [29].

Accessibility in Neuromodulation

Neuromodulation devices also had their accessibility assessed through the application of usability techniques. Glenn has found that most devices incorporate auditory cues, buttons with raised cutouts, speech commands, or other useful features to help people with visual impairments. However, no device has been found that is completely accessible to all users, regardless of visual, auditory and physical limitations [28].

Accessibility in Medical Devices used in Homecare

In addition to medical devices in hospitals, technologies present in the home environment also have accessibility problems, as presented by Blubaugh et al. In his study, the researcher showed that the vast majority of glucometers and blood pressure monitors available on the market have limitations for people with disabilities, especially people with reduced vision [26]. These studies discuss accessibility problems faced with medical devices that compromise the safety of using the technology. Ardehali et al. also studied medical devices used at home, and found in his research that

71% of people with disabilities describe using medical devices as extremely difficult or somewhat difficult [30].

Another study that investigated problems with medical devices in the home was Constance, which explored in detail the types of difficulties experienced by patients with physical/sensory disabilities who use positive airway pressure devices. Problems were reported when performing manual tasks that were difficult for users, such as connecting accessories, changing filters, among others. These demands have contributed to patient frustration and reduced home medical device use [27].

Accessibility in Health Technology Management

In all studies analyzed in the rapid review, accessibility problems were found in medical devices. But there was no evidence of a proposal for a methodology to incorporate accessibility throughout the entire life cycle of technologies, from the development stages to use. Therefore, considering accessibility must be considered at all stages of the life cycle of health technologies, from pre-market to post-market phases, for that, a model was proposed as elucidated in Figure 1. The accessibility should be included in different stages in development of the new technology, and also in planning, acquisition, verification, training, use, decommission and other activities in all life cycles. When applying universal design as a strategy and including people from all ages and abilities throughout the technology lifecycle, from the ideation phase of digital health solutions to development, developers can design solutions with better accessibility. Universal design aims to design products in a safe and autonomous way, in a simple, intuitive way and with equal possibilities of use [5].

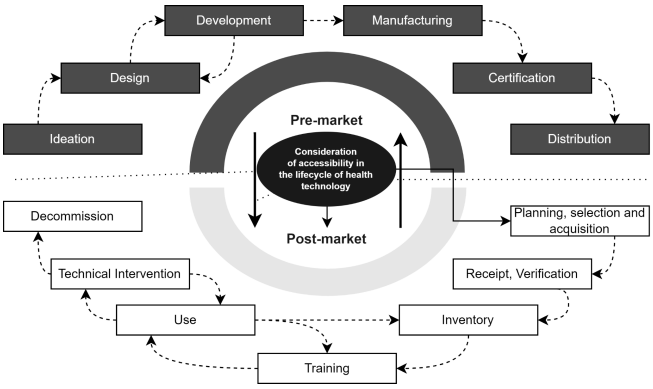


Figure 1. Consideration of accessibility in the lifecycle of technology.

The lack of accessible medical devices is among one of the factors that lead to the disparity in health services available to people with disabilities. It is essential to understand the accessibility and safety barriers present in medical devices used for all types of exams and procedures to meet the diverse population of patients with varying limitations, abilities and disabilities. The model shown in Figure 1 was developed with the application of usability techniques to investigate accessibility problems and improve the usability of medical devices. The application of usability techniques with users at each stage of the life cycle considering population diversity is essential to develop more accessible technologies. The main steps consist of planning the project and defining the objective, studying the technology, choosing the usability techniques to be used, defining the population considering the diversity of users, developing a protocol for applying the technique, applying the protocol with users, analyzing the data, carry out an action plan with preventive strategies and continuously monitor to evaluate effectiveness and seek improvements.

Considering current standards and regulations involving accessibility is a crucial part of the process. Some regulations and documents with accessibility standards for medical devices are listed in Table II.

IV. Discussion

Architectural elements within healthcare facilities represent the most recognized accessibility barriers, but the problems go far beyond stairs and bathrooms. Lack of accessibility in medical equipment is a major concern. More accessible healthcare solutions are critical in promoting equity and achieving health promotion, prevention and security. Consequently, it can help reduce disparity, increase inclusion and make healthcare spaces more equitable.

According to the report of one of the users of the research conducted by Story et al.: *"it takes more than ramps to solve the health care crisis for people with disabilities"* [10]. It is necessary to develop technologies focused on population diversity through the involvement of users from the initial design process of medical equipment. Continuously carrying out training with the entire team and developing standard operating procedures are other strategies to be implemented by Clinical Engineering together with other actors in order to establish a more accessible healthcare environment. In the pre-market stage of medical device development, the lack of inclusion of user diversity in the design and validation of medical devices can result in performance problems of these devices for individuals from certain population profiles, thus perpetuating structural inequalities in medical care. As presented by Jamali, evidence highlights the need to include diverse patient populations in the design and validation of medical devices [43], as biased data used to develop medical technologies is a common root cause of performance variation between racial and ethnic groups [59].

TABLE II. CURRENT REGULATIONS AND GUIDES WITH ACCESSIBILITY STANDARDS FOR MEDICAL DEVICES.

Name	Description
WCAG [32]	Web Content Accessibility Guidelines (WCAG)
ABNT NBR 17060:2022 [5]	Accessibility in mobile device applications - Requirements
ABNT NBR ISO 9241-171:2018 [33]	Ergonomics of Human-System Interaction Part 171: Software Accessibility Guidance
ABNT NBR IEC 60601-1-11:2012 [31]	Electrical medical equipment Part 1-11: General requirements for basic safety and essential performance — Collateral Standard: Requirements for electrical medical equipment and electrical medical systems used in domestic healthcare environments.
ABNT NBR 9050:2021 [38]	Accessibility to buildings, furniture, spaces and urban equipment
Law Nº 13.146/ 2015 [9]	Establishes the Brazilian law on the inclusion of people with disabilities.
Law Nº 10.098/2000 [34]	It establishes general standards and basic criteria for promoting accessibility for people with disabilities or reduced mobility, and provides other measures.
Regulatory Standard NR 17. Ministry of Labour [35]	Brazilian Ergonomics Regulatory Standard.
Guidance & Resources ADA [2]	Americans with Disabilities Act (ADA) regulations. Access to Medical Care for Individuals with Mobility Disabilities
Standards for Accessible Medical Diagnostic Equipment [36]	The Architectural and Transportation Barriers Compliance Board (Access Board or Board) is issuing accessibility standards for medical diagnostic equipment
Enforceable Accessible Medical Equipment Standards [37]	Developed by the National Council on Disability. Enforceable Accessible Medical Equipment Standards: A Necessary Means to Address the Health Care Needs of People with Mobility Disabilities

To consider accessibility in different stages of the technology life cycle is essential. Table III explains the activities carried out in each stage. Interdisciplinary involvement is also important in the process of incorporating new technologies, in order to ensure that the equipment to be incorporated meets the diversity of the population. Medical devices can also be racially or ethnically biased if design flaws lead to performance differences in patients from racial or ethnic minority groups. While these design flaws may be largely unintentional, every effort must be made to identify, mitigate and remove these biases so that they do not contribute to major health disparities in minority groups [59].

To mitigate accessibility problems in medical devices, different areas and professionals must be involved. Some actions and recommendations consist of raising awareness among health professionals about the accessibility problems faced by medical devices, taking population diversity into account in the process of technological development within a living lab ecosystem, and improving the regulatory requirements for devices.

With each innovation, new accessibility problems may arise. As such, it is critical to engage universal design principles from the earliest stages of the manufacturing process to ensure that inclusive devices are designed and accessible to all users, which can ultimately improve device usability, adherence and effectiveness [28]. Several emerging technologies are being increasingly used in

healthcare, such as artificial intelligence, augmented and virtual reality, Internet of Things, blockchain, among others. Inserting accessibility aspects from the beginning of development is crucial to developing accessible solutions. The diffusion of medical devices into Homecare is another challenge. It is necessary to establish and implement measures that aim to assist in the safety and ergonomics of these technologies for the most varied types and profiles of patients, from those with greater technological skills to those with no aptitude at all [31]. It is necessary to establish strategies to guide patients in the use of these technologies and consider the diversity of users and context of use.

The limitations of this work consist of limited use of databases to search for evidence on accessibility in medical equipment, which may lead to the non-consideration of other work that addresses the topic; low number of works analyzing the accessibility of medical equipment considering the users' perspectives.

TABLE III. CURRENT REGULATIONS AND GUIDES WITH ACCESSIBILITY STANDARDS FOR MEDICAL DEVICES.

Life cycle stage	**Main activities**	**Objective to consider accessibility**
Design and development	- Innovation ideation; - Design, prototyping and development; - Compliance with regulations; - Regulations, good manufacturing practices, certification; - Production, distribution, storage, marketing.	- Establishing project goals and requirements considering accessibility based on the problems identified by users; - Testing solutions with the user for validation, usability and accessibility analysis for developing solutions centered on user
Planning and selection	- Health Technology Assessment (HTA) - Analyzing the technologies, infrastructure and human resources to understand the need for incorporation - Checking that the technology has been regularized - Carry out economic analyses, total cost of ownership; - Specifying and select the technology - Purchasing process (bidding if necessary)	- Meeting user needs and considering accessibility when specifying technology, check that technological development is user-centered and based on standards; - Consider accessibility principles and usability techniques to select and to incorporate into Health Technology Assessment;
Installation	- Install the equipment in compliance with the manufacturer's regulations and recommendations	- Evaluate the accessible infrastructure to check the implications for users; - Understand the difficulties faced by users when interacting with the infrastructure;
Training	- Ongoing and periodic training program to ensure that operators are able to carry out their activities; - Drawing up and implementing good practice guidelines for the proper use of health technologies.	- Train users to operate the technology properly; - Train users about accessibility; - Develop training focused on solving problems faced by users; - Develop accessible Good Practice materials for proper use;
Use	- Risk management - Draw up and implement standardized procedures and protocols for the use of technologies - Develop methodologies to ensure technological traceability - Analyze the history of failures and analyze the probable causes - Investigate the adverse events involved.	- Understand the problems of using the technology and understand the impact of accessibility on the occurrence of adverse events. - Analyze the cause of failures incorporated into risk management in order to establish improvement strategies. - Analyze accessibility problems in order to establish specific strategies and improvements in new technological solutions.
Obsolescence, decommissioning and final disposal	- Developing and implementing procedures describing the criteria for decommissioning technology, taking into account the technical, operational, financial or strategic aspects of the establishment. - Execution of the activity by issuing a decommissioning report	- Analyze the effectiveness of using the technology and aspects evolving accessibility; - Evaluate the needs of the accessible technology to ascertain the need for technological replacement; - Researching technological advances that consider accessibility aspects for technologies with better usability.

V. Conclusion and future work

This work highlighted accessibility problems involving medical devices. Through a rapid review of the literature, it was found that most technologies are inaccessible and/or absent within healthcare environments. The fundamentals of accessibility must be incorporated from the beginning of technological development, throughout the other stages of the life cycle of health technologies. This research reinforced the low number of publications involving accessibility assessment in medical devices, and highlights the need to conduct more research incorporating the diversity of user profiles in the development process to make technology management more inclusive and accessible for the entire population.

Due to the reality of the low amount of evidence and research conducted considering accessibility, for future work the Institute of Biomedical Engineering (IEB-UFSC) intends to carry out research carried out with users to highlight accessibility problems in medical equipment inserted in the Living Lab ecosystem, will feature integration with both patients and healthcare professionals, technology manufacturers, clinical engineering, architecture, and other areas and professionals involved. For that, usability techniques will be applied to explore more problems and establish strategies to improve the design of the medical equipment in health. To implement the Living Lab is essential to create an interdisciplinarity and collaborative Health Ecosystem, for the development of accessible and inclusive technologies for all people.

References

[1] M. Brandão and R. Garcia, "Analysis of Accessibility in Medical Devices in Health Technology Management", SMART ACCESSIBILITY 2024 : The Ninth International Conference on Universal Accessibility in the Internet of Things and Smart Environments. Barcelona, Spain. 2024.

[2] Americans with Disabilities Act. ADA: *Access to Medical Care for Individuals with Mobility Disabilities. Guidance & Resources. [Online]*. Available from: https://www.ada.gov/resources/medical-care-mobility/

[3] World Health Organization. WHO: *Global atlas of medical devices: WHO medical devices technical series*. [Online]. Available from: https://iris.who.int/bitstream/handle/10665/255181/9789241512312-eng.pdf

[4] M. Brandão and R. Garcia, "Descriptive analysis of user-centered usability techniques to health technology management," CNIB 2020 National Congress of Biomedical Engineering, no. 43, Mexico, 2020.

[5] Brazilian Association of Technical Standards. ABNT NBR 17060:2022 - Accessibility on mobile devices. Brazil. 2022.

[6] Brazilian Association of Technical Standards. ABNT NBR ISO 9241-11:2021: Ergonomics of human-system interaction Part 11: Usability: Definitions and concepts. Brazil, 2021.

[7] Brazilian Institute of Geography and Statistics (IBGE). *Continuous National Household Sample Survey aimed at people with disabilities.* 2023. [Online]. Available from: https://www.gov.br/mdh/pt-br/assuntos/noticias/2023/julho/brasil-tem-18-6-milhoes-de-pessoas-com-deficiencia-indica-pesquisa-divulgada-pelo-ibge-e-mdhc

[8] World Health Organization. WHO: *Disability, Key facts.* 2023. [Online]. Available from: https://www.who.int/health-topics/disability#tab=tab_1

[9] Law Nº. 13,146, of July 6, 2015. [Online]. Available from: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13146.htm

[10] M. Follette Story, E. Schwier and J. I. Kailes, "Perspectives of patients with disabilities on the accessibility of medical equipment: examination tables, imaging equipment, medical chairs, and weight scales," Disability and health journal, vol. 2, pp. 169–179, 2009, doi:10.1016/j.dhjo.2009.05.003.

[11] O. Vl. Bitkina, H. K. Kim and J. Park," Usability and user experience of medical devices: An overview of the current state, analysis methodologies, and future challenges," International Journal of Industrial Ergonomics, vol. 76, 2020, doi:102932. 10.1016/j.ergon.2020.102932.

[12] Brazilian Association of Technical Standards. ABNT NBR IEC 62366:2016 Health products — Application of usability engineering to health products. Brazil, 2016.

[13] R. Jeffrey and C. Dana, Handbook of Usability Testing: how to plan, design, and conduct effective tests. 2. ed. Indianapolis: Wiley, 2008.

[14] F. E. Ritter, G. D. Baxter and E. F. Churchill, Foundations for Designing User-Centered Systems. London. 2014.

[15] Brazilian Association of Technical Standards. ABNT ISO/TR 16982:2014: Ergonomics of human-system interaction — Usability methods that support the project user-centric. Brazil, 2014

[16] A. C. Tricco, et al., "A scoping review of rapid review methods," BMC Med, vol 13, 2015, doi:10.1186/s12916-015-0465-6

[17] T. J. Lasserson, J. Thomas and J. P. T. Higgins, Chapter 1: Starting a review. In: J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.1 (updated September 2020). Cochrane, 2020.

[18] Ministry of Health. *Methodological guidelines: preparation of a systematic review and meta-analysis of comparative observational studies on risk factors and prognosis.* Department of Science and Technology. Brazil, 2014.

[19] D. Moher, D. Moher, et al., "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," International Journal of Surgery, vol. 8, pp. 336-341,2010, doi:10.1016/j.ijsu.2010.02.007.

[20] M. A. Morris et al., "Use of Accessible Examination Tables in the Primary Care Setting: A Survey of Physical Evaluations and Patient Attitudes," J Gen Intern Med, 2017.

[21] R. Nancy, M. L. Breslin, M. Liang and S. Yee, "Physical accessibility in primary health care settings: Results from California on-site reviews," Disability and Health Journal, vol. 5, pp. 159-67, 2012.

[22] L. I. Iezzoni, et al., "Use of Accessible Weight Scales and Examination Tables/Chairs for Patients with Significant Mobility Limitations by Physicians Nationwide," Joint Commission journal on quality and patient safety, vol. 47, pp. 615–626, 2021, doi:10.1016/j.jcjq.2021.06.005.

[23] N. R. Mudrick, M. L. Breslin, J. Blackwell, X. Wang, and K. A. Nielsen, "Accessible medical diagnostic equipment in primary care: Assessing its geographic distribution for

disability equity," Disability and health journal, vol. 16, pp. 101425, doi:10.1016/j.dhjo.2022.101425.

[24] N. Agaronnik, E. G. Campbell, J. Ressalam, L. I. Iezzoni, "Accessibility of Medical Diagnostic Equipment for Patients With Disability: Observations From Physicians," Archives of physical medicine and rehabilitation, vol. 100, pp. 2032–2038, doi:10.1016/j.apmr.2019.02.007.

[25] M. Lester, D. Eidson, S. Blair, S. Gray, P. Sapp, F. J. Zupancic, B. C. Marshall, and A. Berlinski, "Cystic Fibrosis Foundation Nebulizer and Compressor Accessibility Survey," Respiratory care, vol. 66, pp. 1840–1847, doi:10.4187/respcare.09197.

[26] M. V. and M. M. Uslan, "Accessibility attributes of blood glucose meter and home blood pressure monitor displays for visually impaired persons," Journal of diabetes science and technology, vol. 6, pp. 246–251, doi:10.1177/193229681200600206.

[27] C. H. Fung, U. Igodan, C. Alessi, J. L. Martin, J. M. Dzierzewski, K. Josephson, B. J. Kramer, "Human factors/usability barriers to home medical devices among individuals with disabling conditions: in-depth interviews with positive airway pressure device users," Disabil Health, vol. 8, pp. 86-92, Jan. 2015, doi: 10.1016/j.dhjo.2014.06.002.

[28] B. Glenn, V. Tieppo Francio, B. D. Westerhaus, J. Goree, N. H. Strand, D. Sparks and E. Petersen, E, "Accessibility and Ease of Use in Neuromodulation Devices," Neuromodulation: journal of the International Neuromodulation Society, vol. 27, pp. 584–588, doi:10.1016/j.neurom.2023.03.003.

[29] B. C. Yankaskas, P. Dickens, J. M. Bowling, M. V. Jarman, K. Luken, K. Salisbury, J. Halladay and C. E. Lorenz, "Barriers to adherence to screening mammography among women with disabilities," Am J Public Health, vol. 100, pp. 947-53, May. 2010, doi:10.2105/AJPH.2008.150318.

[30] M. Ardehali; et al, "People With Disabilities & Medical Device Accessibility: Perspectives, Challenges, and the Role of Occupational Therapy Practitioners," The American Journal of Occupational Therapy, 2023, doi:10.5014/ajot.2023.77S2-PO14.

[31] Brazilian Association of Technical Standards. ABNT NBR IEC 60601-1- 11:2012: Electrical medical equipment Part 1-11: General requirements for basic safety and essential performance — Collateral Standard: Requirements for electrical medical equipment and electrical medical systems used in domestic healthcare environments. Brazil, 2012.

[32] Web Content Accessibility Guidelines (WCAG). 2023. [Online]. Available from: https://www.w3.org/TR/WCAG21/

[33] Brazilian Association of Technical Standards. ABNT NBR ISO 9241-171:2018. Ergonomics of Human-System Interaction Part 171: Software Accessibility Guidance.

[34] Brazil Law nº 10.098/2000. General standards and basic criteria for promoting accessibility for people with disabilities or reduced mobility, and provides other measures. 2000.

[35] Ministry of Labor. Ergonomics Regulatory Standards, NR 17. [Online]. Available from: https://www.gov.br/trabalho-e-emprego/pt-br/acesso-a-informacao/participacao-social/conselhos-e-orgaos-colegiados/comissao-tripartite-partitaria-permanente/normas-regulamentadora/normas-regulamentadoras-vigentes/nr-17-atualizada-2023.pdf

[36] Architectural and Transportation Barriers Compliance Board. Standards for Accessible Medical Diagnostic Equipment. [Online]. Available from: https://www.federalregister.gov/documents/2024/07/25/2024-16266/standards-for-accessible-medical-diagnostic-equipment

[37] National Council on Disability. Enforceable Accessible Medical Equipment Standards. [Online]. Available from: https://www.ncd.gov/report/eames-report/

[38] Brazilian Association of Technical Standards. ABNT NBR 9050:2021. Accessibility to buildings, furniture, spaces and urban equipment.

[39] D. M. Burton, M. G. Enigk and J. W. Lilly, "Blood glucose meters and accessibility to blind and visually impaired people," J Diabetes Sci Technol, vol. 6, pp. 242-5, Mar. 2012, doi:10.1177/193229681200600205.

[40] C. Macdonald, H. Lunt, M. Downie and D. Kendall, "How Satisfied Are Patients When Their Choice of Funded Glucose Meter Is Restricted to a Single Brand?" J Diabetes Sci Technol, vol. 11, pp. 1001-1006, Sep. 2017, doi:10.1177/1932296817693016.

[41] H. K. Akturk, J. Snell-Bergeon and V. N. Shah, "Health Care Professionals' Perspectives on Use of Diabetes Technologies for Managing Visually Impaired Patients With Diabetes," J Diabetes Sci Technol, vol. 17, pp. 1610-1613, Nov. 2023, doi:10.1177/19322968221101629.

[42] L. Heinemann, D. Drossel, G. Freckmann and B. Kulzer, "Usability of Medical Devices for Patients With Diabetes Who Are Visually Impaired or Blind," J Diabetes Sci Technol, vol. 10, pp. 1382-1387, Nov. 2016, doi:10.1177/1932296816666536.

[43] H. Jamali, L.T. Castillo, C. C. Morgan, J. Coult, J. L. Muhammad, O. O. Osobamiro, E. C. Parsons and R. Adamson, "Racial Disparity in Oxygen Saturation Measurements by Pulse Oximetry: Evidence and Implications," Ann Am Thorac Soc, vol. 19, pp. 1951-1964, Dec. 2022, doi:10.1513/AnnalsATS.202203-270CME.

[44] E. R. Gottlieb, J. Ziegler, K. Morley, B. Rush and L. A. Celi, "Assessment of Racial and Ethnic Differences in Oxygen Supplementation Among Patients in the Intensive Care Unit," JAMA Intern Med, vol. 182, pp. 849-858, Aug. 2022, doi:10.1001/jamainternmed.2022.2587.

[45] S. E. K. Sudat, P. Wesson, K. F. Rhoads, S. Brown, N. Aboelata, A. R. Pressman, A. Mani and K. M. J. Azar. "Racial Disparities in Pulse Oximeter Device Inaccuracy and Estimated Clinical Impact on COVID-19 Treatment Course," Am J Epidemiol, vol. 192, pp. 703-713, May. 2023, doi:10.1093/aje/kwac164.

[46] Z. Vesoulis, A. Tims, H. Lodhi et al. "Racial discrepancy in pulse oximeter accuracy in preterm infants," J Perinatol, vol. 42, pp. 79–85, 2022, doi:10.1038/s41372-021-01230-3.

[47] N. R. Mudrick, J. Blackwell, M. L. Breslin and X. Wang, "Change Is Slow: Acquisition of Disability-Accessible Medical Diagnostic Equipment in Primary Care Offices over Time. Health Equity," vol. 8, pp. 157-163, Mar. 2024, doi:10.1089/heq.2023.0155.

[48] T. A. Ahmed, N. Bradley and S. Fenesan, "Dental management of patients with sensory impairments," Br Dent J, vol. 233, pp. 627-633, Oct. 2022, doi:10.1038/s41415-022-5085-x.

[49] K. Lakshmi and P. D. Madankumar, "Development of modified dental chair to accomodate both wheelchair bound patients and general population," Disability and Rehabilitation: Assistive Technology, vol. 15, pp. 467–470, 2020, doi:10.1080/17483107.2019.1710775.

[50] M. Nora, B. Alessandra, V. Jean-Noel, M. Martin, R. Jacqueline and B. Christophe, “A scoping review on dental clinic accessibility for people using wheelchairs,” Spec Care Dentist, vol. 41, pp. 329–339, 2021, doi:10.1111/scd.12565.

[51] H. Levy and R. Lena, “Tools and Equipment for Managing Special Care Patients Anywhere, Dental Clinics of North America,” vol. 60, pp. 567-591, 2016, doi:10.1016/j.cden.2016.03.001.

[52] M. Y. Ishaque, S. Rahim and M. H. Hussain, "Factors that limit access to dental care for person with disabilities: Access To Dental Care for Disabled,” Pak Armed Forces Med J, vol. 66, pp. 230-34, Apr. 2016.

[53] F. Rashid-Kandvani, B. Nicolau and C. Bedos, “Access to Dental Services for People Using a Wheelchair,” Am J Public Health, vol. 105, pp. 2312-7, Nov. 2015, doi:10.2105/AJPH.2015.302686.

[54] H. M. Gach, S. L. Mackey, S. E. Hausman, D. R. Jackson, T. L. Benzinger, L. Henke, L. A. Murphy, J. L. Fluchel, B. Cai, J. E. Zoberi, J. Garcia-Ramirez, S. Mutic and J. K. Schwarz, “MRI safety risks in the obese: The case of the disposable lighter stored in the pannus,” Radiol Case Rep, vol. 14, pp. 634-638, Mar. 2019, doi:10.1016/j.radcr.2019.02.023.

[55] E. M. Andresen, J. J. Peterson-Besse, G. L. Krahn, E. S. Walsh, W. Horner-Johnson and L. I. Iezzoni. “Pap, mammography, and clinical breast examination screening among women with disabilities: a systematic review,” Womens Health Issues, vol. 23, pp. 205-14, Jul. 2013, doi:10.1016/j.whi.2013.04.002.

[56] J. R. Pharr, “Accommodations for patients with disabilities in primary care: a mixed methods study of practice administrators,” Glob J Health Sci, vol. 6, pp. 23-32, Oct. 2013, doi:10.5539/gjhs.v6n1p23.

[57] F. R. Andiwijaya, C. Davey, K. Bessame, A. Ndong and H. Kupe, “Disability and Participation in Breast and Cervical Cancer Screening: A Systematic Review and Meta-Analysis,” Int J Environ Res Public Health, vol. 19, pp. 9465, Aug. 2022, doi:10.3390/ijerph19159465.

[58] N. Agaronnik, A. El-Jawahri and L. Iezzoni, “Implications of Physical Access Barriers for Breast Cancer Diagnosis and Treatment in Women with Mobility Disability,” J Disabil Policy Stud, vol. 33, pp. 46-54, Jun. 2022, doi: 10.1177/10442073211010124.

[59] M. W. Sjoding, S. Ansari and T. S. Valley, “Origins of Racial and Ethnic Bias in Pulmonary Technologies,” Annu Rev Med, vol. 74, pp. 401-412, Jan. 2023, doi:10.1146/annurev-med-043021-024004.

[60] E. Arana-Chicas, A. Kioumarsi, A. Carroll-Scott, P. M. Massey, A. C. Klassen and M. Yudell, “Barriers and facilitators to mammography among women with intellectual disabilities: a qualitative approach,” Disabil Soc, vol. 35, pp. 1290-1314, 2020, doi:10.1080/09687599.2019.1680348.

Quality and Governance Framework for the National Telemedicine Network in Greece

Angeliki Katsapi
Euro-Mediterranean Institute of Quality and Safety in Healthcare
Athens, Greece
e-mail: akatsapi@eiqsh.eu

Haralampos Karanikas
Department of Computer Science and Biomedical Informatics
University of Thessaly
Lamia, Greece
e-mail: karanikas@uth.gr

Mariana Tsana
Euro-Mediterranean Institute of Quality and Safety in Healthcare
Athens, Greece
e-mail: mtsana@eiqsh.eu

Fotios Rizos
Euro-Mediterranean Institute of Quality and Safety in Healthcare
Athens, Greece
e-mail: frizos@eiqsh.eu

Vasileios Tsoukas
Department of Computer Science and Biomedical Informatics
University of Thessaly
Lamia, Greece
e-mail: vtsoukas@uth.gr

George Koukoulas
2nd Healthcare Region of Piraeus and Aegean
Piraeus, Greece
e-mail: koukoulas@2dHR.gov.gr

Dimitrios Drakopoulos
Dextera Consulting
Athens, Greece
e-mail:
ddrako@dexteraconsulting.com

Abstract—**This study examines the quality and patient safety dimensions of telehealth with regard to existing standards and regulative provisions aiming to the development of a standard practice framework for the National Network of Telemedicine in Greece. The main purpose of the Greek National Telemedicine Network (EDIT) is to improve healthcare access in Greece, especially on isolated islands and distant mountainous regions. The expansion of EDIT network currently, foresees the establishment of a significant number of new telemedicine stations and the installation of home-care units. This signifies the progression of the system and the growing level of coverage of the population and the provision of services at a larger scale. The present study focuses on the determination of the preconditions, operational rules, elements of the governance framework, and the functional specifications of EDIT as a regulative basis to be established in Greece. Moreover, this work aims to support the implementation of tele-health services in the country by safeguarding all quality aspects of the service including the safety and experience of the user as well as the adequacy of the applied resources. The examined set of prerequisites and quality criteria revealed essential adjustments to the current regulative framework and the ethical code of practice for healthcare professionals in Greece.**

Keywords-e-health; telemedicine; framework; healthcare; Greek national telemedicine network.

I. INTRODUCTION

The application of information technology at the level of health and social care provides nowadays the possibility of comprehensive support and monitoring of both chronic patients and those with low-prevalence diseases, while at the same time promoting the culture and knowledge of prevention and public health. However, financial issues are not the only challenge. Inequalities in access to health resources and structures are evident even among citizens of the same country and health system. In many cases, telehealth and telemedicine services can keep those in need of medical care safely at home and out of hospitals or clinics, providing timely access to diagnosis and treatment, and monitoring chronic problems on a systematic basis [1]. Telehealth does not imply an increase in the quantity of healthcare services offered, but rather the provision of more streamlined and efficient services by healthcare practitioners. Adopting telehealth can present difficulties, but it is undoubtedly achievable [2].

More recently, telehealth has been intensively proposed as a tool to improve the efficiency of health services, as it allows the sharing and coordination of resources that are geographically distant or the redesign of health services to optimize the use and management of available resources (human and logistical).

There is a changing trend in healthcare delivery towards more personalized and patient-focused solutions, through technological advancements, which will provide a significant opportunity to increase healthcare access, particularly in underserved or rural regions, and for individuals who may encounter obstacles in accessing conventional health services.

Responding to these challenges, the Greek Ministry of Health has been investing in the expansion of the National

Telemedicine Network, particularly to extend healthcare coverage of the isolated islands and rural mountainous regions, and other inaccessible areas, fulfilling its constitutional mandate of providing equal healthcare access to all citizens, regardless of their location of residence.

Greece's National Telemedicine Network (EDIT), currently, consists of:

- 66 Patient- Doctor Telemedicine Stations (PDTS) located in hospitals, health centers, and multipurpose regional clinics
- 21 Consultant Telemedicine Stations (CTS) are located in 12 Regional Hospitals and tertiary hospitals within the 2nd Greek Health Region (HR) and the National Emergency Centre (NEC)
- More than 170 Home Care Stations (HCS), which are situated in the homes of in-patients or social care facilities inside the 2nd Greek HR's international boundaries.

Additionally, the existing system will be upgraded in the 2nd HR to include more regional equipment and subscription services. Some of the additions are the following:

- Three hundred and fifty-five new Patient Doctor Telemedicine Stations - PDTS - will be placed in particular Health Facilities nationwide.
- Thirty-five new Telemedicine Consultant Telemedicine - CTS will be placed in designated Health Facilities. CTS stations are categorized based on space availability data and operational requirements of each Health Facility.
- Five Telemedicine Training Stations with CTS and PDTS features will serve as training centers for new system users and will be placed in University Hospitals nationwide.
- Home Monitoring Systems - HCS: 3,000 units with direct communication with the EDIT and related software.
- Three new regional Control Centers and one Command & Control Centre at the Ministry of Health.

Teleconsultation services in Greece during the period 2016 to 2023 showed a notable increase in mental health services, and more specifically, telepsychiatry sessions were the most common type of teleconsultation, followed by telepsychiatry for children, diabetology-related consultations and teleconsultations for chronic disease management [1].

In addition, it is proven that telemedicine services in Greece are progressing and offer a valuable chance to enhance healthcare accessibility, especially in underserved or isolated regions, and for individuals facing barriers to traditional health services [2]. An important prerequisite for EDIT to meet its purpose and operational goals is the adoption of appropriate rules and conditions to achieve quality results and desirable clinical outcomes.

In the aftermath of the COVID-19 pandemic, the World Health Organization (WHO) published a study, with the aim of understanding the evolution of digital health, including the physical and technical characteristics of the infrastructure that supports it, its promotion and utilization, and the barriers that may hinder its widespread adoption [3].

The study concludes that in response to the outbreak of the pandemic, an increasing number of countries are developing organized and systematic telemedicine services. Simultaneously, they are proceeding with the implementation of regulatory interventions with relevant legislation or strategies aiming to create conditions for the sustainable and quality provision of wide-ranging telehealth services. Sixty percent (60%) of Member States stated that their telemedicine services have improved due to the pandemic, while 59% of Member States have issued new relevant legislation, strategy, policy or guidelines to support the provision of telehealth services [3]. The area that requires more focus is the evaluation of the services provided in terms of efficiency criteria and the systematic planning of resources and financing of telemedicine services to ensure their uninterrupted and sustainable operation [3]. The contribution of this study is to examine the quality and patient safety dimensions of telemedicine by analyzing existing standards and regulations integrated into a practice framework that meets quality, operational, and user safety objectives in a unified approach. The remainder of this study is organized as follows. Section II presents the methodology for developing the operational quality requirements for the National Telemedicine Network in Greece. The Code of Practice is presented as a baseline in accordance with the multidimensional model for telehealth. In addition, relevant international standards are used to select complementary specifications to cover all functional, technical, resource-related, and procedural areas. In section III, the main results of the previous analysis are elaborated, concluding with the proposed quality framework with the requested criteria and conditions to be applied for the deployment of EDIT. Finally, section IV discusses digital health initiatives and challenges, while section V concludes the study with future steps and objectives.

II. METHODOLOGY

The study carried out by the scientific team and the collaborators of the University had as a main goal to formulate a proposal for the minimum requirements and specifications that the regulatory framework in Greece should adopt, with the aim of applying quality and safety criteria to the expanded operation of the National Telemedicine Network in Greece.

To develop the appropriate requirements, the existing framework of the Code of Good Practice for telehealth services at the European and at international level was reviewed, and all the individual dimensions of quality and safety for health services for telemedicine as defined in the literature were analyzed. On this basis, relevant standards published by the International Organization for Standardization (ISO) and international healthcare accreditation bodies were examined for all quality aspects and functional dimensions through a scientific review. The unique perspective of this study is its integrated approach. While the reference framework examined (international standards,

regulatory requirements, ethical codes) focuses on specific areas of telemedicine, the aim was to go beyond the provisions of the codes of practice to cover safety, person-centeredness, effectiveness, resource competence, and ethical dimensions in a unified manner. In addition, governance aspects for the sustainable development and operation of EDIT were considered, and key roles and responsibilities across the network were incorporated into the proposed framework. Based on this research, the specifications and operational goals developed as the essential prerequisites at an organizational-technical-functional level. The relevant provisions also included requirements concerning the human factor both in terms of the competence of the healthcare professionals providing telehealth services and in relation to the engagement of the beneficiaries (citizens, patients, caregivers). The synthesis of all the above-mentioned provisions was conducted by using the multidimensional model of telehealth [4].

A. Multidimensional Model

Telehealth is supported by a complex business operation involving a set of multi-layered socio-political, economic, organizational, professional, cultural, human, legal, technological and strategic factors. It is therefore very important that all these factors are considered collectively when planning, implementing, developing and evaluating telehealth services, which requires numerous changes and transformations at the micro, meso- and macro-level of the services provided. This study was based on this combination of factors to determine the minimum operational requirements that will frame efficient, effective, value-adding, sustainable, and secure telehealth services of the national network [5].

The study focused on analyzing a multidimensional model to gain a deeper comprehension of the various ways in which telehealth services can manifest in terms of complexity.

The multidimensional model that was studied serves a better understanding of the complexity of the different forms that the use cases of telehealth services can take [4][6]. The reference domains of the multidimensional model are provided in Table 1.

TABLE 1. MULTIDIMENSIONAL MODEL - REFERENCE DOMAINS

Technology	The transition from a physical to a virtual environment, as new challenges or risks usually appear, pre-requests significant changes in the related operations, through application of rigid processes imposed by technology, ergonomics and design or new ways of virtual communication that differ from interpersonal contact.
Human factors	The issue of human-technology interaction raises cognitive and human concerns for implications in terms of cognition, habits, behaviors, memory, mental and cognitive components, psychomotor factors, and individual psychosocial and cultural characteristics. Technology is a set of artifacts that must consider the peculiarities and characteristics of the individual user.
Organization	Health care organizations are complex social systems, with heterogeneity and diversity of individual and group cultures, dynamics, interests, and behaviors. Telehealth services, as a socio-technical objective, could cause a re-definition of balances, workflows and powers, thus creating conflicts of professional and organizational jurisdiction.
Fiscal /policy & regulatory framework	The health sector is usually governed by a strict regulatory framework. The evolution of standards and certification requirements or providers' obligations (e.g., quality, safety - security and privacy), financial and technical policies related to telehealth can create a multitude of unforeseeable consequences that demand a prompt reengineering of services and procedures.

B. Code of Practice, the Existing Provisions

The main precondition for the successful development of telehealth is trust by the health professionals and the service users, and trust is cultivated ultimately by setting standards and external controlling procedures. In the study, the European Code of Good Practice for Tele-Health services, which is a fundamental regulatory document, was examined as a basic reference framework of service requirements.

The European Code also prefaced the International Code of Practice for Telehealth Services, which also offers a standardized approach and a guide for the conformity assessment and the certification of services. By the application of these two documents, a quality benchmark can be created according to which telehealth services (including tele-care) can be assessed [7]. The content of the European and International Code of Good Practice for Tele-Health services is presented in Figure 2 [8].

C. International Standards for Quality Assurance Leading to Certification or Accreditation of Telemedicine Services

Standards and guidelines aim to spread good practices and guarantee a certain level of requirements in the use of telehealth solutions. As in any context of services provided, in telehealth services, it is very important to review already established specifications, conditions and operational objectives to adequately ensure the basic attributes that safeguard quality as well as ethical and safety principles in the provision of services [9][11][12][13][14][15][16][17].

Furthermore, the technical requirements deriving from the relevant standards aim to ensure interoperability between different devices, units, providers' entities and health systems, and thus, it is fundamental to create conditions whereby applications are compatible with other systems.

Quality assurance in health care can be verified and validated by obtaining certification and / or accreditation as a strong indication of an organization's commitment to high-level quality criteria. Figure 1 displays the multi-dimension concept of healthcare access,

D. Special ISO Standard for Telemedicine Services

For the operation of telehealth services, the recently published standard ISO 13131:2021 provides specific quality criteria and control points for the implementation of telehealth services, integrating risk management objectives and procedures as well as quality elements in the form of service design guidelines. The adoption and implementation of the standard aims to optimize the provided services through continuous improvement of applied procedures, standardization of communication aspects, coordination of resources, and clarification of accountabilities, ultimately benefiting both healthcare providers and patients [9].

This standard covers issues regarding:

- Management of processes related to the quality assurance of telehealth services
- Design and implementation of strategic and operational processes related to regulations, best practices and guidelines
- Healthcare procedures involving the beneficiary/ patients
- Management of financial resources for the provision of telehealth services
- Secure management of the information that is circulated and used in the context of telehealth services

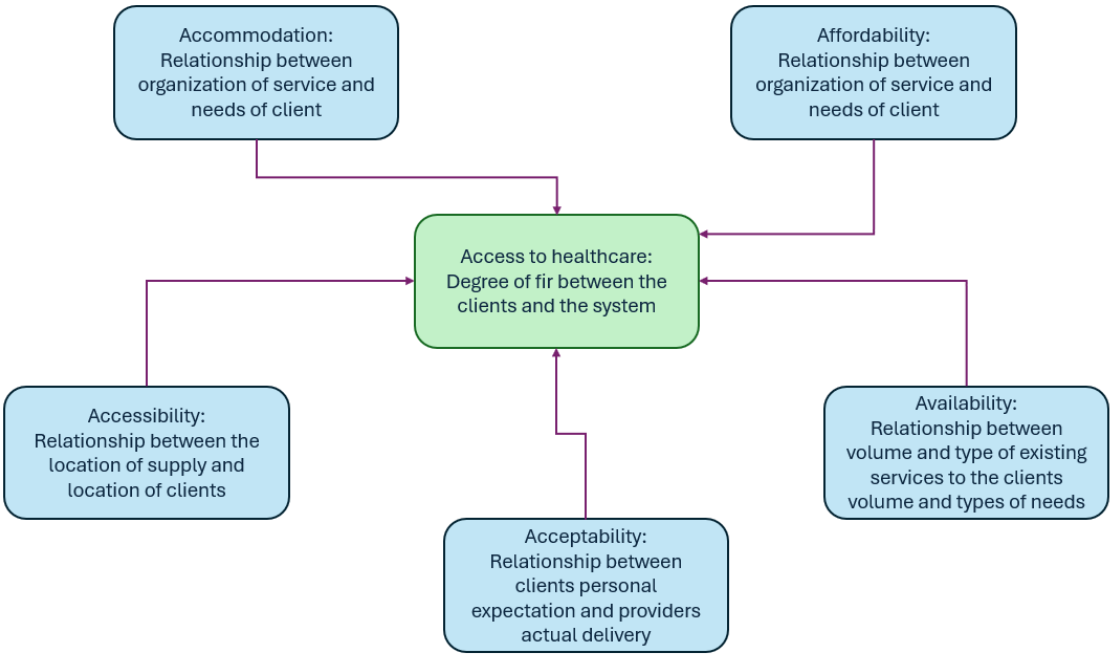


Figure 1. The multi-dimensional concept of health care access

.

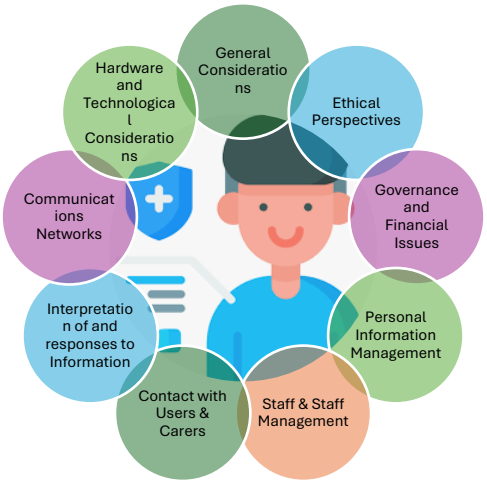


Figure 2. International & European Code of Good Practice for Telehealth Services – Contents

Processes related to the design and provision of human resources, infrastructure facilities, and technological resources for use by telehealth services. Additionally, there are accreditation bodies whose quality standards and criteria are developed to control and evaluate telehealth services aspects.

The technical adequacy and acceptance of these standards on an international scale are ensured through their validation and accreditation by the international organization International Society for Quality in Health Care External Evaluation Association (ISQua EEA) ISQua EEA provides third-party external assessment services to health and social care external assessment organizations and standards development bodies worldwide [10]. An overview of the key health-related accreditation bodies is provided in Table 2.

TABLE 2. KEY HEALTH-RELATED ACCREDITATION BODIES

✓ Joint Commission International (JCI)
✓ Temos International Healthcare Accreditation
✓ National Committee for Quality Assurance (NCQA)
✓ Accreditation Canada
✓ Global Healthcare Accreditation
✓ Utilization Review Accreditation Commission (URAC)

After studying the accreditation programs implemented by healthcare providers including the provision of telehealth, it is concluded that the adaptation of an integrated model of compliance assessment may have a direct impact to human resources engagement and to the upgrade of the quality level and reliability of the services pursuing better performance and clinical efficiency in telemedicine services.

III. RESULTS

As a result of the processing and alignment of requirements and specifications with the services dimensions as indicated by the codes of practice for telemedicine, the minimum requested criteria and conditions have been developed as a framework proposal for adaptation regarding the operation of EDIT including the governance model to be applied for the network.

A. Proposed Framework Requirements for Quality and Safety Assurance of Telehealth Services

i. Licensing procedures to authorize telehealth providers: Healthcare providers who simultaneously provide telemedicine services, regardless of the context in which they operate, should be fully licensed in accordance with relevant administrative, legislative, and regulatory requirements. They are also required to implement the specifications' framework and ensure comprehensive documentation and proof of its implementation.

ii. Quality Assurance System for Telehealth Services (SSP T-Y): in this section, the provisions of a Quality Management System (QMS) to be developed by the provider are described, covering the entire scope, context, purpose, and objectives of the offered services. The outline of the applied processes and protocols and the method for the evaluation of outcomes should be determined. Furthermore, entities responsible for any part of the services, including third parties (health care units, supporting organizations, manufacturers, suppliers, and health service insurance organizations), must be well defined, and relevant authorizations should be addressed.

iii. Risk Management System: The provider, in addition to the standardization of the applied procedures, should have in place a risks' identification and risks' mitigation process towards the achievement of the system objectives, through which the evaluation of the external and internal factors of the organization will be carried out, followed by the prioritization of risks that can prevent each organization from achieving its objectives.

iv. Resources Management System: The provider should ensure quality, suitability, continuity, reliability, effectiveness, availability, safety, sustainability aspects and usability of the infrastructure and equipment, proper support of the remote services, adequacy of devices and technology used for the provision of telehealth services.

v. Special Quality Assurance Issues

- Available Resources: In this section, special provisions have been included for mainly interconnection and telecommunications requirements, which should always be satisfied for each category of telehealth service, namely:
  - Methods of communication and applications
  - Availability of infrastructure and technologies

- Connectivity for interactive meetings
- Systems' interoperability

- Acquiring new skills and competence of healthcare professionals: The minimum required criteria regarding the formal qualifications, necessary knowledge, skills and specialized certifications of healthcare professionals should be determined so that they can provide telemedicine services. Among the most important competence elements to be certified is knowledge of basic aspects and guidelines for:
  - the provision, documentation and reporting of telehealth services (use of the digital systems applied),
  - the available options, telehealth protocols and prescribing requirements that apply both domestically and in the context of cross-border care,
  - risks management and limitations of telehealth services,
  - benefits and limitations of telehealth services.
- Ethical and Consensus Issues: Issues of securing the patient such as his identification, information, informed consent and participation in decision-making for care, but also the respect for the confidentiality and privacy of the circulating information are some key points that are specialized in this point of the study.
- Aspects that must also be covered are (i) the use of appropriate means of communication and applications that have the appropriate authentication, confidentiality and security parameters necessary for proper use, (ii) the availability of infrastructure and technologies using up-to-date security software, (iii) the connectivity for real-time interactive meetings based on bandwidth and frame rate, (iv) interoperability with EHR systems to sustain continuous service provision and integrated care.

B. Governance Model

The governance framework designed for the National Network of Telemedicine serves as a guiding map for all accountabilities and functions that govern the network to ensure alignment with regulatory requirements, compliance with ethical and technical standards, and alignment with stakeholder expectations, while promoting innovation and excellence in the practice of telemedicine. A brief description of the basic principles of governance is provided in Table 3.

TABLE 3. BASIC PRINCIPLES OF GOVERNANCE

Accountability: Creating mechanisms for assigning responsibilities.
Transparency: Enhancing transparency in decision-making processes, policies and actions related to telehealth governance.
Equity: Ensuring equitable access to telehealth services and adequacy of resources for the population, regardless of geographic location, socioeconomic status, or demographic characteristics.
Risk Management: Identifying, assessing and mitigating the risks associated with telehealth services, including technical, operational, legal and ethical risks.
Collaboration and coordination: Strengthen collaboration and coordination among telehealth stakeholders at local, regional, national and international levels.
Quality assurance: The establishment of operational standards, guidelines and quality assurance mechanisms is the basis for ensuring the application of all critical principles for telehealth services.
Regulatory Compliance: Compliance with applicable laws, regulations and standards governing telehealth operation, data privacy, security and interoperability.
Sustainability: Ensuring a sustainable governance framework for telehealth services that can adapt to the changing needs of healthcare, the healthcare ecosystem and technological developments.
Ethical issues: Operating according to established principles and values in telehealth governance, including respect for patient autonomy, privacy, confidentiality and informed consent.
Legal Framework: A strong legal framework that provides clear guidelines and rules governing telehealth practices.
Data Governance: Development of comprehensive data governance policies and procedures that will govern the collection, storage, use, management and protection of health data that are circulated in telehealth systems.
Interoperability Standards: Adoption of interoperability standards and protocols for the seamless access, interconnection and exchange of health information across different telehealth platforms, systems and electronic health record systems Health Record).
Training and Education: Provide training and education programs to build capacity and proficiency in telehealth standards, technology use, and workflow integration.
Technological Proficiency and Innovation: Invest in a robust, scalable technology infrastructure to support telehealth services.
Evaluation and Research: Conducting systematic evaluation and research reports

In the proposed governance model, the senior management (Ministry of Health) provides the strategic direction while the operational management is carried out by the Coordination Directorate with the support of 3 new departments: (i) Business Operations, (ii) Technological Infrastructure and Innovation and (iii) Department of

Communication, Publicity and Digital Media, with responsibilities of coordination, guidance and support of the Operational Centers (servicing the telemedicine stations) in order to meet the needs of smooth daily operation through technological excellence.

Operational Centers are created at each Health Region, which is also responsible for the training and certification of new users, development of partnerships, and supervision and control of the system for routing technical requests to treating gaps and failures of the system.

Ministry of Health regional policies and guidelines for the implementation and operation of telehealth should align with regional healthcare priorities and public health goals. An important part of the strategic planning is the allocation of the corresponding resources.

A Coordinating Committee serves as a high-level advisory and decision-making body, responsible for the supervision of the overall direction and implementation of the national telehealth program. The regional coordination committees are intended to facilitate cooperation between regional bodies involved in the implementation and operation of telehealth and to monitor and evaluate the performance and impact of telehealth services at the local level.

A quality committee is responsible for overseeing quality assurance and improvement efforts within the national telehealth program. It focuses on ensuring that services comply with established standards of care, safety, and effectiveness and that continuous quality improvement processes are in place to improve service delivery and patient outcomes.

The technology standards committee focuses on defining technology standards and protocols related to telehealth technology infrastructure, interoperability, and security. The telehealth user advisory committee represents the voice of patients and community members in telehealth program design, implementation, and evaluation.

C. *Proposed Code of Ethics for the Provision of Telehealth Services*

The lack of a universally accepted code of ethics creates challenges regarding the quality and ethics of service delivery, ensuring at the same time the protection of patients' rights and the professional integrity of providers. Table 4 describes the proposed code of ethics for the provision of telehealth services.

TABLE 4. PROPOSED CODE OF ETHICS FOR THE PROVISION OF TELEHEALTH SERVICES

1.	Tele-health services refer to all versions of interaction and communication between healthcare professionals and the user- patient that take place remotely, as well as the provision of tele-health services as defined internationally/ in a national context.
2.	The doctor / healthcare professional must ensure that the beneficiary of the services (citizen-patient) meets predefined inclusion criteria for the requested service (as determined by the relevant telemedicine protocols).
3.	The doctor / healthcare professional must ensure the informed consent and consensus of the patient / citizen to use telehealth services.
4.	Every type of telecommunication in the provision of tele health services must respect the healthcare professional-patient relationship and individual needs ensuring mutual trust, evidence-based decision making and practice, patient autonomy, privacy and confidentiality of communications.
5.	In any type of telecommunication, the identification method of both the doctor/healthcare professional and the patient/ service user must be ensured.
6.	The content and outcome of each session must be recorded in the clinical file of the patient/ service user.
7.	When the patient participates in a telehealth session in the form of remote consultation, it should be carried out under conditions comparable to a face-to-face visit and the availability of the necessary information should be ensured.
8.	The patient has the right to access evidence of the health professional's competence to accept the provided telehealth services.
9.	The doctor/healthcare professional who provides tele-consultation services to another health professional may provide scientific opinion, recommendation or clinical decisions only if he considers that the exchanged information is sufficient and relevant. In cases where a doctor asks for the opinion of his colleague, he is responsible for the recommendations provided to the patient in the case the recommendations issued remotely by the other doctor.
10.	The doctor/ healthcare professional must recommend the clinical assessment in physical presence, if not possible to obtain the patient's consent after being informed about the implementation of the tele-health service or when it is impossible to implement the tele-session according to the leges- artis.
11.	The doctor / healthcare professional must ensure that security measures are implemented to protect the medical / health record.
12.	The doctor/health professional and the patient have the freedom and complete independence to decide whether to use or refuse the telehealth service application.
13.	The doctor/healthcare professional must ensure that the training and competence of other collaborators involved in the transmission or delivery of data is sufficient.
14.	The doctor must ensure that he has public liability insurance for the use of telemedicine services.

15. In the event of a breach of confidentiality of which the doctor/health professional becomes aware, the doctor must directly inform the patient.

IV. Discussion

Digital Health initiatives such as telehealth, mobile health, and clinical decision-support systems may provide alternative solutions to accessible care, continued surveillance, risk mitigation, clinical outcomes' monitoring, and containment of the disease.

The development of an integral and sustainable national network for remote access to healthcare services requests a thorough study of all the essential components of the implemented systems and a governance structure to supervise and support the continuous improvement of safe and clinically effective operations. A huge challenge for both patients/ service users and telehealth providers is how to determine the reliability, appropriateness and compliance of the applied systems with functional requirements and technical standards.

While Digital health tools are part of the overall frame of Health Technology Assessment, there is a definite need to organize resources and evaluation programs with regard to their compliance against public health requirements and quality criteria and based on their impact on clinical service and the level at which they meet health needs.

V. Conclusion

National networks for telehealth services should be governed by clear policies and strategies at the health system level. These policies should define the role of telehealth in the delivery of healthcare and determine clear structures and service requirements while essential implementation acts are required to satisfy quality criteria and functional specifications that must be addressed by a suitable and properly applied regulatory framework.

At a next phase, the assessment exercise for the National Network of Telemedicine in Greece should be based on a well-structured system of key performance indicators reflecting compliance with the defined standard practice framework and the impact and outcomes to the service users.

References

[1] H. Karanikas, V. Tsoukas, D. Drakopoulos, G. Koukoulas, A. Katsapi, F. Rizos, "Assessing Greek National Telemedicine Network," The Sixteenth International Conference on eHealth Telemedicine, and Social Medicine (eTELEMED 2024), IARIA, May 2024, pp. 43-49, ISSN: 2308-4359, ISBN: 978-1-68558-167-1

[2] GE HealthCare, "Telehealth Best Practices: A Guide for Physicians," 2022. [Online]. Available: https://www.volusonclub.net/empowered-womens-health/telehealth-best-practices-a-guide-for-physicians/ [Accessed: 23 November 2024].

[3] WHO, "The ongoing journey to commitment and transformation: digital health in the WHO European Region," 2023. [Online]. Available: https://www.who.int/europe/publications/m/item/digital-health-in-the-who-european-region-the-ongoing-journey-to-commitment-and-transformation. [Accessed 29 November 2024].

[4] H. Alami, M.-P. Gagnon, and J.-P. Fortin, "Some Multidimensional Unintended Consequences of Telehealth Utilization: A Multi-Project Evaluation Synthesis," Int J Health Policy Manag, vol. 8, no. 6, pp. 337–352, Jun. 2019, doi: 10.15171/ijhpm.2019.12.

[5] Pan American Health Organization, "Framework for the Implementation of a Telemedicine Service," 2017. [Online] Available: https://iris.paho.org/handle/10665.2/28414. [Accessed 29 November 2024].

[6] P. Ouma, P. M. Macharia, E. Okiro, and V. Alegana, "Methods of Measuring Spatial Accessibility to Health Care in Uganda," in Practicing Health Geography: The African Context, P. T. Makanga, Ed., Cham: Springer International Publishing, 2021, pp. 77–90. doi: 10.1007/978-3-030-63471-1_6.

[7] European Commission, "Code of Practice sets benchmark for telehealth services," 2013. [Online]. Available: https://digital-strategy.ec.europa.eu/en/news/code-practice-sets-benchmark-telehealth-services. [Accessed: 29 November 2024].

[8] Telehealth Quality Group EEIG, "International Code of Practice for Telehealth Services 2018/19 v2," 2018. [Online]. Available: https://www.isfteh.org/files/work_groups/2018-19-INTERNATIONAL-TELEHEALTH-CODE-OF-PRACTICE.pdf [Accessed: 29 November 2024].

[9] ISO, "ISO 13131:2021 Health informatics - Telehealth services - Quality planning guidelines," 2021. [Online]. Available: https://www.iso.org/standard/75962.html. [Accessed 29 November 2024].

[10] International Society for Quality in Health Care External Evaluation Association, "Trusted Accreditation for Health & Social Care Evaluators," 2024. [Online]. Available: https://ieea.ch/. [Accessed 29 November 2024]

[11] European Standards, "BS EN 15224:2016," 2017. [Online]. Available: https://www.en-standard.eu/bs-en-15224-2016-quality-management-systems-en-iso-9001-2015-for-healthcare/ [Accessed 29 November 2024].

[12] ISO, "ISO 7101:2023 Healthcare organization management - Management systems for quality in healthcare organizations – Requirements," 2023. [Online]. Available: https://www.iso.org/standard/81647.html. [Accessed 29 November 2024].

[13] ISO, "ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection - Information security management systems – Requirements," 2022. [Online]. Available: https://www.iso.org/standard/27001. [Accessed 29 November 2024].

[14] ISO, "ISO 16527:2023 Health informatics - HL7 Personal Health Record System Functional Model, Release 2 (PHR-S FM)," 2023. [Online]. Available: https://www.iso.org/standard/84665.html. [Accessed 29 November 2024]

[15] ISO, "ISO/TR 9143:2023 Health informatics, Sex and gender in electronic health records," 2023. [Online]. Available: https://www.iso.org/standard/83431.html. [Accessed 29 November 2024].

[16] ISO, "ISO 18104:2023 Health informatics, Categorial structures for representation of nursing practice in terminological systems," 2023. [Online]. Available: https://www.iso.org/standard/81132.html. [Accessed 29 November 2024)

[17] ISO, "ISO/TS 17975:2022 Health informatics - Principles and data requirements for consent in the collection, use or disclosure of personal health information," 2022. [Online]. Available: https://www.iso.org/standard/78395.html. [Accessed 29 November 2024