ACHI 2025

The Eighteenth International Conference on Advances in Computer-Human Interactions

ISBN: 978-1-68558-268-5

May 18 - 22, 2025

Nice, France

ACHI 2025 Editors

Weizhi Meng, Lancaster University, UK

ACHI 2025

Forward

The Eighteenth International Conference on Advances in Computer-Human Interactions (ACHI 2025), held between May 18th, 2025, and May 22nd, 2025, in Nice, France, was a result of a paradigm shift in the most recent achievements and future trends in human interactions with increasingly complex systems. Adaptive and knowledge-based user interfaces, universal accessibility, human-robot interaction, agent-driven human computer interaction, and sharable mobile devices are a few of these trends. ACHI 2025 also brought a suite of specific domain applications, such as e-learning, social, medicine, education, and engineering.

We take here the opportunity to warmly thank all the members of the ACHI 2025 technical program committee, as well as all the reviewers. The creation of such a high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to ACHI 2025. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the ACHI 2025 organizing committee for their help in handling the logistics of this event.

We hope that ACHI 2025 was a successful international forum for the exchange of ideas and results between academia and industry for the promotion of progress in the area of human-computer interactions.

ACHI 2025 Chairs

ACHI 2025 Steering Committee
Sibylle Kunz, IU Internationale Hochschule, Germany
Lasse Berntzen, University of South-Eastern Norway, Norway
Weizhi Meng, Lancaster University, UK
Flaminia Luccio, University of Venice, Italy
Abdul Khalique, Liverpool John Moores University, UK

ACHI 2025 Publicity Chairs
José Miguel Jiménez, Universitat Politècnica de València, Spain
Francisco Javier Díaz Blasco, Universitat Politècnica de València, Spain
Ali Ahmad, Universitat Politècnica de València, Spain
Sandra Viciano Tudela, Universitat Politècnica de València, Spain
Laura Garcia, Universidad Politécnica de Cartagena, Spain

ACHI 2025
Committee

ACHI 2025 Steering Committee

Sibylle Kunz, IU Internationale Hochschule, Germany
Lasse Berntzen, University of South-Eastern Norway, Norway
Weizhi Meng, Lancaster University, UK
Flaminia Luccio, University of Venice, Italy
Abdul Khalique, Liverpool John Moores University, UK

ACHI 2025 Publicity Chairs

José Miguel Jiménez, Universitat Politècnica de València, Spain
Francisco Javier Díaz Blasco, Universitat Politècnica de València, Spain
Ali Ahmad, Universitat Politècnica de València, Spain
Sandra Viciano Tudela, Universitat Politècnica de València, Spain
Laura Garcia, Universidad Politécnica de Cartagena, Spain

ACHI 2025 Technical Program Committee

Mark Abdollahian, Claremont Graduate University, USA
Mostafa Alani, Tuskegee University, USA
Marran Aldossari, University of North Carolina at Charlotte, USA
Obead Alhadreti, Umm Al-Qura University, Al-Qunfudah, Saudi Arabia
Asam Almohamed, University of Kerbala, Iraq
Mehdi Ammi, Univ. Paris 8, France
Anmol Anubhai, Amazon, USA
Prima Oky Dicky Ardiansyah, Iwate Prefectural University, Japan
Mohd Ashraf Bin Ahmad, University Malaysia Pahang, Malaysia
Charles Averill, University of Texas at Dallas, USA
Snježana Babić, Juraj Dobrila University, Croatia
Matthias Baldauf, OST - Eastern Switzerland University of Applied Sciences, Switzerland
Giulio Barbero, Leiden University (Leiden Institute of Advanced Computer Science), The Netherlands
Catalin-Mihai Barbu, University of Duisburg-Essen, Germany
Yacine Bellik, IUT d'Orsay | Université Paris-Saclay, France
Lasse Berntzen, University of South-Eastern Norway, Norway
Ganesh D. Bhutkar, Vishwakarma Institute of Technology (VIT), Pune, India
Cezary Biele, National Information Processing Institute, Poland
Christos J. Bouras, University of Patras, Greece
Christian Bourret, UPEM - Université Paris-Est Marne-la-Vallée, France
Sabrina Bouzidi-Hassini, Ecole nationale Supérieure d'Informatique (ESI), Algeria
James Braman, The Community College of Baltimore County, USA
Justin Brooks, University of Maryland Baltimore County / D-Prime LLC, USA
Pradeep Buddharaju, University of Houston - Clear Lake, USA
Paolo Burelli, IT University of Copenhagen, Denmark
Idoko John Bush, Near East University, Cyprus

Athina Papadopoulou, Massachusetts Institute of Technology (MIT), USA
Evangelos Papadopoulos, National Technical University of Athens, Greece
Vida Pashaei, University of Arizona, USA
Dennis Paulino, INESC TEC / University of Trás-os-Montes e Alto Douro, Portugal
Freddy Alberto Paz Espinoza, Pontificia Universidad Católica del Perú, Peru
Gerald Penn, University of Toronto, Canada
Jorge Henrique Piazentin Ono, New York University - Tandon School of Engineering, USA
Ana C. Pires, Universidade de Lisboa, Portugal
Jorge Luis Pérez Medina, Universidad de Las Américas, Ecuador
Brian Pickering, IT Innovation Centre - University of Southampton, UK
Thomas M. Prinz, Friedrich Schiller University Jena, Germany
Annu Sible Prabhakar, University of Cincinnati, USA
Mike Preuss, Leiden University, Netherlands
Namrata Primlani, Northumbria University, UK
Marina Puyuelo Cazorla, Universitat Politècnica de València, Spain
Yuanyuan (Heather) Qian, Carleton University in Ottawa, Canada
Claudia Quaresma, Universidade NOVA de Lisboa, Portugal
Aiswarya R., Tata Consultancy Services, India
Neha Rani, University of Florida, USA
Mariusz Rawski, Warsaw University of Technology, Poland
Carsten Röcker, inIT - Institute Industrial IT / TH OWL University of Applied Sciences and Arts, Germany
Joni Salminen, Qatar Computing Research Institute, Qatar
Sandra Sanchez-Gordon, Escuela Politécnica Nacional, Ecuador
Antonio-José Sánchez-Salmerón, Instituto de Automática e Informática Industrial - Universitat Politecnica de Valencia, Spain
Paulus Insap Santosa, Universitas Gadjah Mada - Yogyakarta, Indonezia
Markus Santoso, University of Florida, USA
Diana Saplacan, University of Oslo, Norway
Hélène Sauzéon, Centre Inria Bordeaux, France
Daniel Schneider, Federal University of Rio de Janeiro, Brazil
Trenton Schulz, Norwegian Computing Center, Norway
Kamran Sedig, Western University, Ontario, Canada
Sylvain Senecal, HEC Montreal, Canada
Fereshteh Shahmiri, Georgia Tech, USA
Yuhki Shiraishi, Tsukuba University of Technology, Japan
Zdzisław Sroczyński, Silesian University of Technology, Gliwice, Poland
Ben Steichen, California State Polytechnic University, Pomona, USA
Han Su, RA - MIT, USA
Federico Tajariol, University Bourgogne Franche-Comté, France
Sheng Tan, Trinity University, Texas, USA
Cagri Tanriover, Intel Corporation (Intel Labs), USA
Ranjeet Tayi, User Experience - Informatica, San Francisco, USA
Masashi Toda, Kumamoto University, Japan
Milka Trajkova, Indiana University, Indianapolis, USA
David Unbehaun, University of Siegen, Germany
Simona Vasilache, University of Tsukuba, Japan
Katia Vega, University of California, Davis, USA
Vignesh Velmurugan, University of Hertfordshire, UK

Nishant Vishwamitra, University at Buffalo, USA
Konstantinos Votis, Information Technologies Institute | Centre for Research and Technology Hellas, Greece
Lin Wang, U.S. Census Bureau, USA
Pinhao Wang, Zhejiang University - College of Computer Science and Technology, Hangzhou, China
Gloria Washington, Howard University, USA
Andreas Wendemuth, Otto-von-Guericke University, Germany
Zhanwei Wu, Shanghai Jiao Tong University, China
Shuping Xiong, KAIST, South Korea
Tong Xue, Beijing Film Academy, China
Bo Yang, The University of Tokyo, Japan
Rui Yang, Xi'an Jiaotong-Liverpool University, China
Ye Zhu, Cleveland State University, USA

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Psychological Modelling and Action Recognition for Boxing Performance

Vishal Ravishankar
Computer Science Dept.
PES University
Bengaluru, India
email: v1sh4lrr@gmail.com

Itish Raj Shukla
Computer Science Dept
PES University
Bengaluru, India
email: itish2606@gmail.com

Nishat-E-Afroz M
Computer Science Dept
PES University
Bengaluru, India
email: nishatm1717@gmail.com

Anirudh Bhatta J A
Computer Science Dept.
PES University
Bengaluru, India
email: anirudh.bhatta@gmail.com

Surbhi Choudhary
Computer Science Dept.
PES University
Bengaluru, India
email: surbhi.choudhary@pes.edu

***Abstract*— Psychological dynamics in boxing remain largely unaddressed during real-time performance evaluation, despite their significant impact on decision-making and reaction time. Traditional analysis tools focus predominantly on physical performance, overlooking the internal cognitive states of athletes. This gap limits coaches' and psychologists' ability to intervene strategically during or after a match. The Boxing Psychological State Tracking System is a novel framework designed to analyze a boxer's psychological state throughout a fight by correlating physical performance indicators with inferred cognitive conditions. By leveraging action recognition and explainable Artificial Intelligence (AI), the system evaluates events such as knockdowns, strike patterns, and defensive behavior to infer mental states like fatigue, disengagement, and stress. Previous approaches to action recognition in sports have either ignored psychological interpretation or lacked transparency in decision-making, which this work addresses. Our solution involves a lightweight, vision-based system combining You Only Look Once, version 8 - nano variant (YOLOv8n) and SHapley Additive exPlanations (SHAP) to map strike behavior to inferred cognitive states. The system integrates these insights with a visual analytics interface, enabling coaches, sports psychologists, and athletes to understand and improve performance through a deeper awareness of psychological dynamics. Evaluation of the system demonstrates its potential to reveal meaningful psychological patterns using just two high-frequency strikes, providing a practical and explainable method for mental state assessment in combat sports.**

Keywords*— *Sports psychology; Action recognition; Boxing behavior; Convolution neural network.

I. INTRODUCTION

In competitive boxing, performance depends not only on physical ability but also on psychological resilience. Mental states like fatigue, disengagement, and stress can significantly affect a boxer's decision-making and reaction times, yet tracking these psychological states during a fight remains a difficult challenge. Current tools focus mainly on physical metrics, leaving a gap in real-time psychological analysis. This paper introduces the Boxing Psychological State Tracking System, a vision-based tool that infers a boxer's cognitive state by analyzing performance indicators such as strike patterns, knockdowns, and movement changes.

The primary difficulty in this domain lies in the absence of real-time tools that can interpret mental states from visible behavioral cues, complicating the ability to intervene during a match. By using computer vision models like YOLOv8 and Region-based Convolutional Neural Network (R-CNN) for action detection, and SHAP for interpretability, the system highlights behavioral cues linked to psychological states. Our research is driven by the challenge of mapping psychological states directly to physical indicators observed in real-time boxing matches. Our work investigates three main research questions: Can psychological conditions be inferred from observable fight behavior? Which physical cues best indicate mental fatigue or lapses? And how can these insights be effectively visualized to assist athletes and coaches? The system features a user-friendly interface that displays real-time plots, SHAP overlays, and psychological state indicators, making mental patterns easy to understand and act upon. The explicit purpose of this article is to present a framework that bridges the gap between action recognition and cognitive state inference, providing a novel approach to analyzing performance in combat sports. By bridging action recognition with cognitive modeling, our work contributes a novel and explainable approach to performance analysis in combat sports. One limitation of this approach is that it currently focuses on a limited set of strike types, and further work is needed to account for a broader range of actions and defensive maneuvers.

The paper is organized as follows: Section II reviews related work in the field of action recognition and psychological state tracking. Section III outlines the methodology behind the Boxing Psychological State Tracking System, including the dataset used and the models employed. Section V presents the results, followed by a discussion of the findings in Section VI. Section VII concludes the paper and Section VIII suggests directions for future work.

II. RELATED WORK

This section reviews prior research in the field. While no existing systems have specifically focused on tracking psychological states during boxing matches, our approach leverages advancements in action recognition. Models like YOLO, R-CNN, and other convolutional neural networks have been successfully used to detect and classify actions in video data. We build on these techniques, adapting them to analyze boxing-specific movements and perform psychological analysis in real time.

 ISBN: 978-1-68558-268-5

A. Boxing Behaviour Recognition based on Artificial Intelligence CNN with sports psychology assistant

The main intent of the research was to develop a mechanism to recognize boxing form with the help of AI-Convolutional Neural Network (CNN) and also to examine how the state of mind of athletes affects the accuracy and the effectiveness of behavior recognition. The research deployed a mixture of tools like psychological assessment survey and AI technologies to have an insight into fighters' psychological profiling and for the construction of boxing action classification and recognition algorithms. They developed the model using Bidirectional Encoder Representations from Transformers **(**BERT) fusion 3D-Residual Networks (ResNet) architecture, which provided the emotion conveying info along with action features. The model suggested in this research was very effective compared to the traditional models, the loss value, along with the accuracy and F1 values were improved, where the accuracy reached 96.86% [1].

B. Deep Learning for Micro-Expression Recognition: A Survey

In this report, the mission is to conduct a complete survey of Deep Learning (DL) approaches toward Micro-expression Recognition (MER). The purpose of the article is to form a taxonomy for the field that illustrates all aspects of MER based on deep learning. It will be integrated with existing linguistic datasets and deep learning methodology, and also compare performances of the key DL methods. The paper makes a quick review of the related obstacles, e.g., difficulties in data gathering and annotation, data scarcity, and the dynamical MEs which are subtle, spontaneous and extremely fast. The bottom of the manuscript has a set of DL approaches which have been recommended to solve these issues and improve accuracy of MER. The paper proceeds further, indicating that deep learning has shown clearly the superb performances in MER too. The article notably mentions the still remaining issues against which of the MER mechanisms should be evaluated [2].

C. A Video Based Human Detection and Activity Recognition –A Deep Learning Approach

The goal is to achieve a match of the accuracy of existing systems used for Human Activity Recognition (HAR) systems that helped in the recognition of different human actions in different video clips. The proposed model is the CNN and it is a high-performance architecture, which is designed for pixels as inputs and aids in image recognition and other processing. The models are based on the CNN architecture and are trained on a set of human action's videos. Then, their performance is assessed using the standard set of videos. The suggested model surpasses the current status-quo shown by the above state-of-the-art on the HAR data set. The model shows a good performance and yields a high accuracy of 79.33% for the frame based and of 84.4% for the image-based measurement [3].

D. Factors affecting concentration of attention in boxing athletes in combat situations

This research focuses specifically on the determining the parts that play a key role in getting boxers to possess the right amount of concentration during the boxing matches. This research will talk about the self-assessment of athletes and coaches on attention skill to understand features of lack of focus and the role external factors (such as loud noises or crowd reaction) in recognition among competitors. As the analytical approach, there is the practice of interviewing boxers and the acquisition of their self-understanding data on how they perceive their concentration indicators. Training session beginnings and competitive team play allow for optimization processes which includes human oversight. This test is used to examine situations, which are responsible for disturbing players' attentional processes. On the other hand, coaches' assessments on the concentration of the boxers is a rated external metric that is used to assess attention. Data is gathered and analysed by using statistical methods including descriptive statistics analysis and non-parametric variance study methods to interpret it. The study revealed the role of certain factors that might have a negative effect on serious arousal of boxers, such as the perception of their opponent's superiority and the criticism by their trainer during the critical situations. These results, on the other hand, mean the high significance of the Self and external triggers bringing athletes to be either attentive or not [4].

E. Relationship between selected psychological variables among trainees of combat sports

The task of this study was to look into how the selected psychological parameters could influence the behaviour of the Trainees in combat sports- Aggression, Sports competition anxiety, and Sports Achievement motivation. Sports sciences laboratory conducted study with 10 male athletes in each of the three sports (Boxing, Wrestling, and Judo). The test subjects were exposed to Buss Perry Aggression Questionnaire (BPAQ), Sports Competition Anxiety Test (SCAT) and Sports Achievement Motivation Test (SAMT) which measure aggression, sports competition anxiety, and motivation towards sports achievement accordingly [5]. It was analysed using the tools of Descriptive Statistics and the procedure of Pearson Product-moment correlation coefficient. The outcome indicates that Sports Competition Anxiety is related with Sports Achievement Motivation (r = -0.45; p 0.05).

F. A Unified Approach to Interpreting Model Predictions

The SHAP framework of interpreting complex machine learning models addresses the critical challenge that several existing methods share in a unified theoretical foundation rooted in Shapley values from game theory. Model interpretability is critical in building trust, improving models, and understanding processes generally, especially in light of the increasing use of so-called "black-box" models such as ensembles and deep networks. SHAP defines a class of additive feature attribution methods that satisfy desirable properties such as local accuracy, missingness, and consistency that many prior approaches lack. By unifying six existing techniques, such as Local Interpretable Model-agnostic Explanations (LIME) and Deep Learning Important FeaTures (DeepLIFT), SHAP reveals their commonalities and resolves their limitations. New algorithms, such as Kernel SHAP and Deep SHAP, are proposed to estimate feature importances efficiently, which makes it suitable for both model-agnostic and model-specific contexts. SHAP outperforms existing methods in providing more accurate, computationally efficient, and human-intuitive explanations, as shown through experiments on decision trees, deep

networks, and user studies. Despite the computational costs for large datasets, SHAP's progress represents a significant step toward transparent and trustworthy AI, with ongoing work focused on further enhancing its scalability and explanatory power [6].

III. METHOD

In this section, we describe the architecture and components of the proposed system, which integrates action classification, model explainability, and a user-facing visualization interface.

A. Action Classification Model Building

The first step involves selecting a model architecture capable of accurately analyzing fight sequences in images and video frames. Given the nature of the task—detecting and classifying rapid, overlapping movements—CNNs are a natural fit due to their strong performance in image-based tasks [3].

We initially considered R-CNN and Faster R-CNN [3], known for their ability to focus on specific regions of interest within an image using region proposals. However, these models are computationally expensive and not optimized for real-time analysis.

Instead, we selected YOLOv8n [7] due to its speed and accuracy in multi-object detection. YOLO is a one-stage detector that performs object localization and classification in a single forward pass, making it suitable for real-time video analysis. It is particularly effective at identifying multiple actions in complex scenes—such as various strikes and movements in a boxing match.

The model was fine-tuned on our custom dataset with annotated classes representing boxing actions such as jab, hook, uppercut, knockdown, and defensive movements. We adjusted hyperparameters such as network depth, learning rate, and number of epochs to optimize performance for our task.

B. Use of SHAP

To improve the transparency and trustworthiness of our model, we integrated SHAP [6] a game-theoretic approach to explain the output of machine learning models.

SHAP [6] assigns importance values (SHAP values) to different input features—in this case, regions of the image—indicating how much each part contributed to the model's final prediction. This allows researchers and behavioral analysts to understand not just what the model predicted, but why it made that decision.

The visual output from SHAP highlights the regions most influential in classifying specific actions, offering insight into both model performance and potential areas for improvement.

C. Visualization Interface

To make the results accessible and actionable, we developed a user-friendly visualization interface designed to support both analytical and interpretive tasks. The interface presents real-time plots of detected actions across a timeline [10], allowing users to track momentum shifts and behavioral patterns throughout a fight. It also includes overlay visualizations of SHAP explanations directly on the video frames, highlighting the specific regions that influenced the model's decisions. Additionally, the interface displays inferred psychological state indicators, derived from a combination of recognized actions and contextual cues. The system is also capable of identifying targeted attentional breaks—such as pauses in activity, delayed reactions, or shifts in defensive posture—by analyzing temporal patterns in the action data. Specifically, these moments are detected through noticeable fluctuations in the ratio of strikes over time [4], where a sudden drop or irregularity may indicate a lapse in focus or engagement. Such patterns are then mapped to behavioral scales representing cognitive disengagement, mental fatigue, or stress.

IV. BOXING DATASET

A. Data Collection

To design a boxing dataset that captured all the essential features, around 20 videos were downloaded from YouTube, with a broad cross section of weight categories, video quality, and sources to assure generalizability from the model. The videos were split into a series of frames and then reviewed frame by frame to find critical moments in the fighting processes. Many features related to the fights were captured and analyzed. LabelImg was the annotation tool that was used to draw bounding boxes around the region of interest and the annotations were stored in Extensible Markup Language (XML) format for better parsing and accessibility. All boxing videos, images, and annotations saved in Google Drive were further accessed from Google Colab to develop the model.

B. Dataset Preparation

LabelImg is the annotation tool that was used to label the frames individually. Repetitive frames were discarded to avoid overfitting. Different aspects of the fight were captured to help the model understand the nature of the fight better.

The labels are shown in Table I.

TABLE I. DATASET LABELS

ID	*Object/Action*	*Label*
0-1	Boxers	boxer1/boxer2
2-3	Successful Straight Punch	successfulstraightpunchboxer1/ successfulstraightpunchboxer2
4-5	Unsuccessful Straight Punch	unsuccessfulstraightpunchboxer1/ unsuccessfulstraightpunchboxer2
6-7	Successful Hook	successfulhookboxer1/ successfulhookboxer2
8-9	Unsuccessful Hook	unsuccessfulhookboxer1/ unsuccessfulhookboxer2
10-11	Cuts	boxer1cut/boxer2cut
12-13	Knockdowns	boxer1knockdown/ boxer2knockdown

These labels capture all the crucial aspects of the fight. The images were then flipped to increase the strength of the dataset and its ability to classify the frames. After this the images went through the standard process of resizing for uniformity and then normalizing. Normalizing would normalize the RGB values to a value between 0 and 1. This concludes the dataset preparation.

(a) successful straight punch (b) boxer knockdown

Fig. 1. Examples of different boxing actions.

As shown in Figure 1, key boxing actions such as successful straight punches and knockdowns are illustrated, which form the basis of our dataset labeling process.

V. RESULTS

In this section, we evaluate the performance of the model in action recognition.

Our boxing model provides a solution to understand the impact of mental health on a boxer's performance. By analyzing scientific data and body movements during a fight, the system is able to provide a better view of an athlete's condition during each round. Below, we discuss the main features, functions, and user feedback that demonstrate the effectiveness of this system. The system can also detect important physical signs that affect an athlete's mental and physical health, such as cuts. The recommendations suggest that this feature is particularly useful for performance. To help users understand the interaction between the mind and body, our system uses interactive visualization to present information. These features help coaches and athletes identify key points during competition where mental state changes significantly, allowing for a clear understanding of how this change impacts well-being. For example, users may notice an increase in their stress levels after a knockdown, or a boost in confidence after a successful strike such as a hook or a straight punch. By providing a clear annotation for each psychometric test, SHAP can help users identify which factors make up the majority of the sample output, thus providing confidence in the system's results. This feature will be of particular benefit to model developers who want to understand the logic behind each psychological test to make more informed decisions. The user-friendly photo upload function simplifies the analysis process, allowing instructors to seamlessly upload combat photos. Once downloaded, the system uses TensorFlow and OpenCV to identify and isolate frames which bring value or where changes are evident. This frame removal helps users focus on the most important issues, keeping analysis fast and relevant. User feedback suggests that these features increase usability, as they provide access to good, detailed information without long waiting times.

A. Action Classification Model Output

Figure 2 shows the model's output on various images from different fights. We can see that the model is accurately able to classify all the different labels it has been trained on with high confidence.

Fig. 2. Image Action Classification Output.

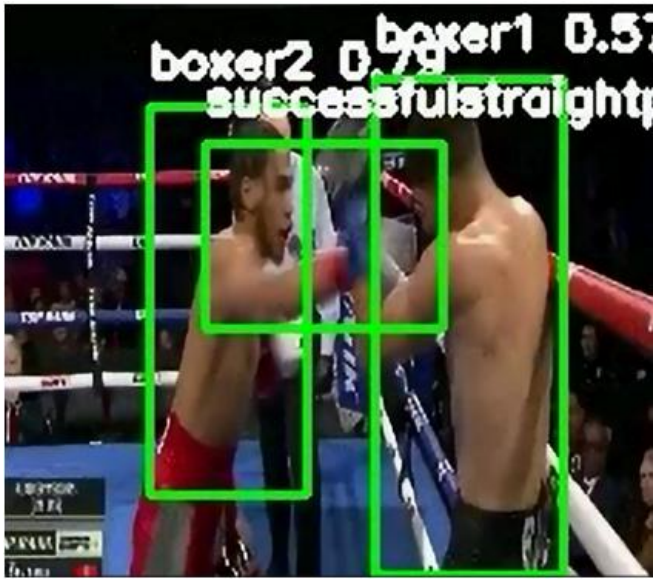

Fig. 3. Video Action Classification Output.

Figure 3 shows the model's prediction on a fight video. Our model is able to perform near real time classification on any fight video with high accuracy. Figure 3 depicts boxer 1 and boxer 2 where boxer 2 throws a successful straight punch, our model was able to correctly classify this action.

B. Action Classification Model Evaluation

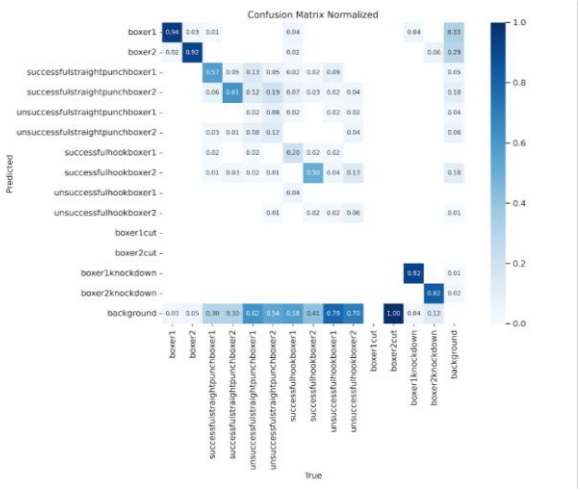

Fig. 4. Confusion Matrix.

Figure 4 shows the confusion matrix which can be used to evaluate the model's performance. Each cell in the matrix shows the proportion of samples classified into a particular category relative to the true category. Darker shades correspond to higher proportions which represents better classification performance of the model for that specific true-predicted pair. Classes like boxer1(true) vs. boxer1 (predicted) has a high proportion of correct classifications (~0.94), indicating strong performance of model for this class. Whereas, some confusion occurs between classifying the type of punch, straight or hook which indicates the need for a rich featured dataset. Specific events like boxer knockdown and cuts are classified well. This shows the overall performance of model on different classes.

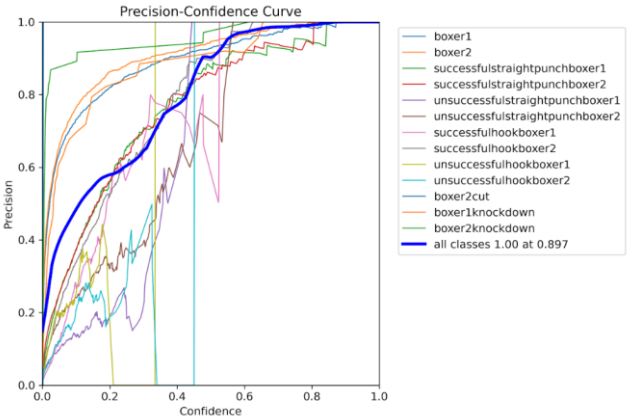

Fig. 5. Precision Curve.

Figure 5 shows the Precision-Confidence Curve which can be used to visualize the relationship between the model's confidence in its predicted and the precision achieved for various classes. Each curve shows how the precision changes as confidence threshold increases for a particular class. The blue line indicates that the model achieves perfect precision i.e., 1.0 for predictions when the confidence threshold is set to 0.897.

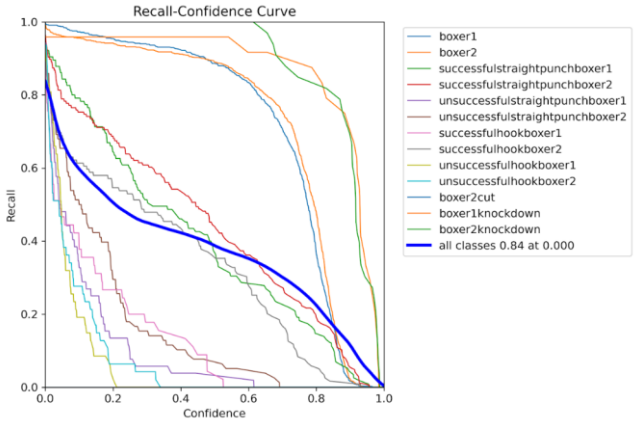

Fig. 6. Recall Curve.

Figure 6 shows the Recall curve for different classes. Recall measures the ability of the model to find all relevant instances in the dataset. It is calculated as the number of true positives divided by the sum of true positives and false negatives. A higher recall means that the model is capturing a larger portion of relevant cases. The curve starts at a higher recall levels when the confidence threshold is low which means that the model is more inclusive but less accurate. As the confidence threshold increases, the recall decreases.This indicates that the model is becoming more selective, thus more confident but potentially missing some true positive instances.

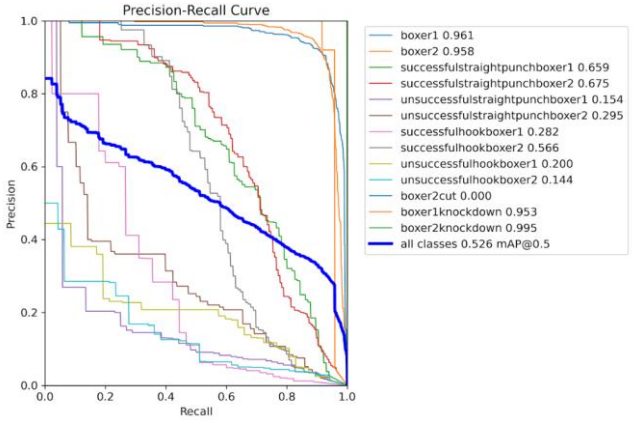

Fig. 7. Precision Recall Curve.

Figure 7 shows a Precision-Recall (PR) curve which is another way to evaluate the performance of the model. This helps in understanding the trade-off between precision and recall for each class at different thresholds of classification decision. At higher precision levels recall is generally low indicating the model being very selective, making fewer predictions but those predictions more likely to be correct. In the right side of the graph, recall is high and precision ends to be low which indicates that the model identifies most of the positive cases but also makes false positive errors. Understanding PR curve is crucial where the cost of false positives are different from tat of false negatives. When it comes to action classification, it is better to miss a positive case than incorrectly labelling a negative case as positive. This curve helps in finding the right balance based on specific needs.

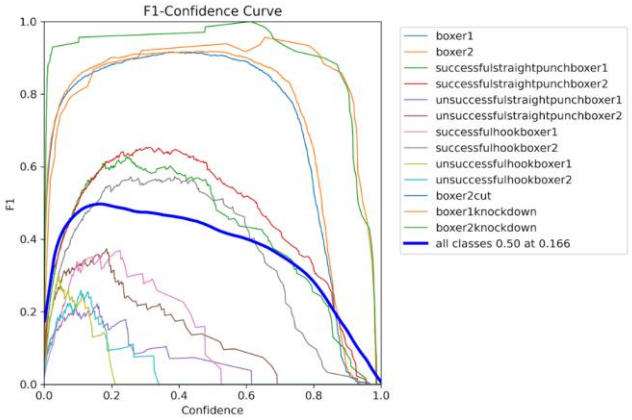

Fig. 8. F1 Curve.

Figure 8 shows F1 curve for different classes, this curve combines precision and recall into a single metric. Classes like boxer 1 and boxer 2 maintain high F1 scores across confidence levels which indicates strong performance in these classifications. The blue line indicates that the average F1 score across all classes at a confidence threshold of 0.166 is 0.50.

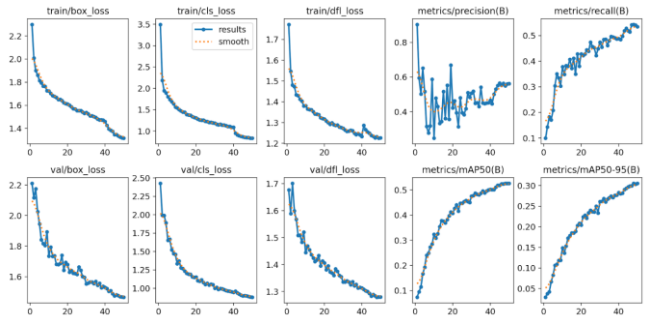

Fig. 9. YOLO Model Results.

Figure 9 provides a comprehensive overview of performance metrics over epochs during the training and validation phase of the model. The overall trend of decreasing loss and increasing precision and recall is a positive sign that the model is learning effectively from the training data and improving its prediction capabilities as training progresses.

C. *Psychological Analysis of Boxer's Performance*

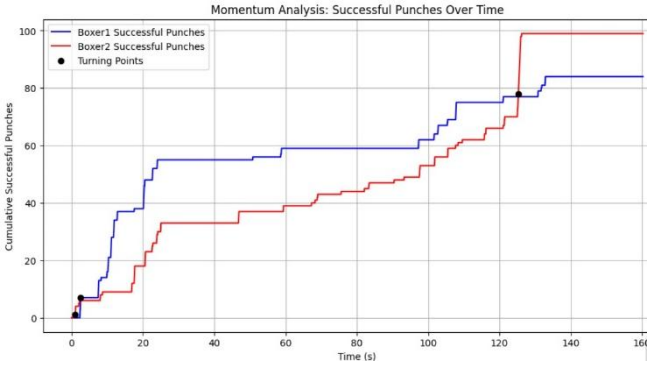

Fig. 10. Momentum Analysis: Turning Points.

Figure 10 shows the cumulative number of successful punches landed by two boxers in the match. The blue line tracks cumulative successful punches of boxer 1 and red line tracks cumulative successful punches of boxer 2, with the steepness reflecting the scoring pace. Steeper sections denote more rapid scoring of points. Black squares mark specific moments in the match labeled as "Turning Points." These indicate moments where the momentum of the match shifts significantly. This can be due to a critical strike, a tactical change, or other significant event that affects the dynamics of the contest. By comparing the trajectories of two lines we can infer that boxer 1 starts strong, gaining an early lead. However, boxer 2 increases his rate of successful punch as the match progresses eventually surpassing boxer 1. The black squares or turning points are crucial for coaches as they highlight moments when potential strategic changes might have influenced the match's outcome. Coaches can make use of this data to improve training regimens, focusing on shifting or maintaining momentum at critical stages of the match.

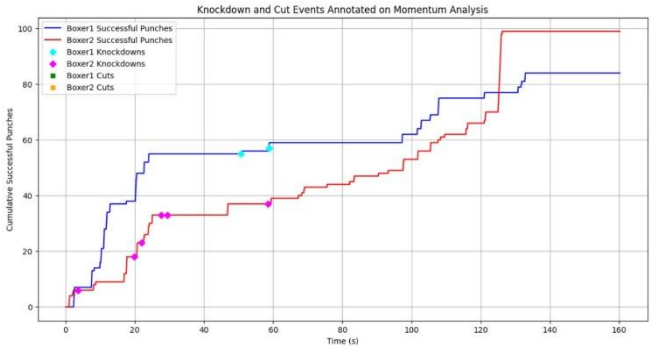

Fig. 11. Momentum Analysis: Knockdowns and Cuts Annotated.

Figure 11 expands on the earlier momentum analysis by including events like knockdown and cuts for understanding the impact of significant events on the momentum of a match. Initially, both boxer 1 and boxer 2 accumulate points at similar rates with boxer 1 taking a slight lead. The first significant event in the above graph is the knockdown of boxer 2. Despite this, boxer 2 begins to accumulate punches at higher rate. Shortly after boxer 1 experiences knockdowns around 50 to 60-second mark which marks a clear momentum shift in boxer 2's favor. At the 80-second mark another key turning point occurs where boxer 2 shows a surge in successful punches, significantly outpacing boxer 1. This can be correlated with strategic adjustments or a decline in boxer 1's defense. The match ends with boxer 2 having a clear lead in successful punches, possibly reflecting better stamina or strategy execution. This graph offers a detailed view of how events like knockdowns and cuts can impact the flow of match which can be used to train the boxers to enhance their performance in the future matches.

Figure 12 represents the distribution of action between boxer 1 and boxer 2 across different rounds of a boxing match. Each pie chart depicts the percentage of total actions taken by boxer 1 and boxer 2. For trainers and coaches, these insights are valuable for assessing the resilience and recovery of a boxer after being knocked down. This can also help in strategizing training to enhance endurance and tactical responses to such events in the upcoming matches. Overall, these enhanced momentum analysis graphs serve as a tool for in-depth review and strategic planning in sports, particularly in boxing.

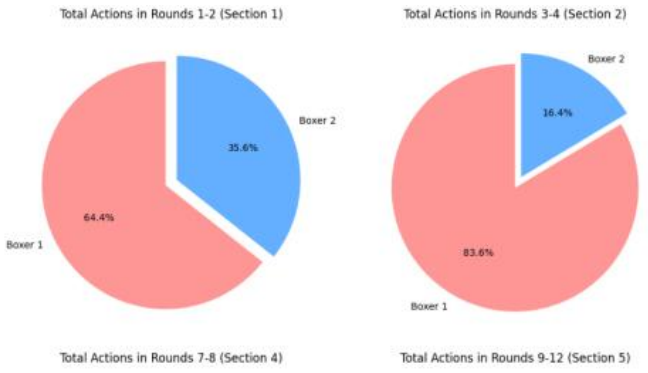

Fig. 12. Total Actions in Sections.

VI. Discussion

As admitted, the current model is limited to recognizing only two primary strike types—straight punches and hooks. This design choice was intentional and guided by the objectives of the system. For our current goal of inferring psychological states, focusing on high-frequency, high-impact strikes like the straight and hook is sufficient. These are not only the most commonly thrown punches but also indicative of strategic intent, energy levels, and mental engagement, aligning with previous findings on how psychological factors shape athletic focus and execution in combat sports [4], [5], [13]. Thus, variations in their frequency, timing, and accuracy serve as strong proxies for cognitive states such as fatigue, disengagement, or stress.

Previous efforts in the domain of sports psychology have mostly focused on either physical performance metrics or indirect psychological assessments. Existing tools, such as traditional biometric sensors or motion tracking systems, fail to offer real-time, interpretable insights into the cognitive states of athletes. Moreover, while some systems incorporate physical actions like punch tracking, they often lack the ability to connect these actions to mental states, leaving a gap in understanding how psychological factors influence physical performance. Our system addresses this gap by leveraging real-time action recognition and explainable AI techniques [1], [6], [8] to correlate strike patterns with inferred psychological conditions, offering a unique, practical approach for performance optimization.

However, while our system offers significant advancements, it is not without limitations. The current approach is based on a narrow set of strike types, and there is a clear need to expand it to account for a broader range of actions, including defensive movements, counter-strikes, and combinations. Additionally, the system's reliance on visual data means it could be enriched with additional inputs such as physiological signals (e.g., heart rate, galvanic skin response) for a more comprehensive understanding of a boxer's mental state. These features are on our wish-list for future iterations of the system. Moreover, we plan to further refine our movement recognition model to improve accuracy and robustness in different lighting and environmental conditions. Another goal is to incorporate more advanced psychological tests that could enhance the system's ability to identify specific mental conditions like anxiety or stress, which may not be fully captured by physical performance metrics alone.

The purpose of our contribution is to offer a practical, real-time tool for tracking mental states in combat sports, empowering athletes and coaches to make data-driven decisions. The system provides a unique combination of performance analysis and mental health monitoring, which can guide training regimens, optimize decision-making in competition, and contribute to athlete well-being by identifying early signs of mental fatigue or stress. By offering clear visualizations and intuitive feedback, the system bridges the gap between physical performance and psychological analysis, supporting athletes in maintaining peak performance while safeguarding their mental health.

Looking forward, we envision several areas of improvement. First, the addition of more complex strike patterns, defensive actions, and physiological signals will help enhance the system's accuracy and completeness. Second, improving the user interface through user feedback and further research will make the system even more intuitive and applicable across different sports. Finally, further usability surveys and structured testing with domain experts will ensure that the system continues to meet the practical needs of coaches, athletes, and sports psychologists. Ultimately, our goal is to create a tool that not only supports performance enhancement but also prioritizes the mental health and well-being of athletes, ensuring their longevity in the sport.

VII. Conclusion And Future Work

In this work, we set out to bridge the gap between physical action recognition and cognitive state inference in competitive boxing. We developed the Boxing Psychological State Tracking System, a vision-based framework that uses deep learning models like YOLOv8 and R-CNN to detect key actions, and SHAP to provide interpretable insights into inferred psychological states such as fatigue, confidence, and stress. By analyzing behavioral cues like strike patterns, movement changes, and knockdowns, our system offers an accessible interface that visualizes mental dynamics in real time, supporting coaches and athletes in making more informed decisions during performance analysis. The system demonstrates high accuracy in classifying trained strike types and offers a novel, explainable perspective on mental health in combat sports.

Looking ahead, we aim to enhance the system's capabilities and user experience in several key areas. Future work will consider a broader range of psychological factors beyond confidence and anxiety, such as opponent strength and team identity. For example, incorporating data on relative rankings or prior performance against specific opponents may provide deeper insights into mental state fluctuations. Additionally, we plan to expand the model's action recognition capabilities to include more strike types such as jabs, enabling broader applicability. The system could also evolve to offer predictive insights based on historical data, guiding both training and tactical preparation. Furthermore, we envision extending the framework to other combat sports like fencing and mixed martial arts, where only the labeling methodology would need adjustment. Finally, to improve generalizability and robustness, we intend to expand the dataset beyond the current set of 20 fights, ensuring more diverse and representative coverage of combat scenarios.

REFERENCES

[1] Y. Kong and Z. Duan, "Boxing behaviour recognition based on artificial intelligence convolutional neural network with sports psychology assistant," *Scientific Reports*, vol. 14, no. 1, pp. 7640, Apr. 2024.

[2] Y. Li, J. Wei, Y. Liu, J. Kauttonen, and G. Zhao, "Deep Learning for Micro-Expression Recognition: A Survey," IEEE Trans. Affect. Comput., vol. 13, no. 4, pp. 2028–2046, Oct.-Dec. 2022.

[3] M. Dhar and B. Chakraborty, "A video based human detection and activity recognition–a deep learning approach," Turk. J. Comput. Math. Educ., vol. 11, no. 1, pp. 551–559, 2022.

[4] M. C. Hernández, Y. S. Prieto, J. D. García, and M. S. Rodríguez, "Factors affecting concentration of attention in boxing athletes in combat situations," Rev. PODIUM, vol. 15, no. 1, pp. 5–21, Apr. 2020.

[5] A. Sharma and P. Purashwani, "Relationship between selected psychological variables among trainees of combat sports," J. Sports Sci. Nutr., vol. 2, no. 1, pp. 1–3, 2021.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS'17), Red Hook, NY, USA, 2017, pp. 4768–4777.

[7] P. Stefański, J. Kozak, and T. Jach, "The problem of detecting boxers in the boxing ring," in *Recent Challenges in Intelligent Information and Database Systems*, E. Szczerbicki, K. Wojtkiewicz, S. V. Nguyen, M. Pietranik, and M. Krótkiewicz, Eds., *Communications in Computer and Information Science*, vol. 1716, Singapore: Springer, 2022, pp. 546–556.

[8] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, "Explanation of machine learning models using shapley additive explanation and application for real data in hospital," *Comput. Methods Programs Biomed.*, vol. 214, pp. 106584, 2022.

[9] V. Hudovernik and D. Skocaj, "Video-based detection of combat positions and automatic scoring in jiu-jitsu," in Proc. 5th Int. ACM Workshop Multimedia Content Analysis Sports, Oct. 2022, pp. 55–63.

[10] J. Brindha and G. Nallavan, "A wearable biometric performance measurement system for boxing - A survey," in 2022 IEEE World Conf. Appl. Intell. Comput. (AIC), Sonbhadra, India, 2022, pp. 536–540.

[11] R. Merlo et al., "Profiling the physical performance of young boxers with unsupervised machine learning: A cross-sectional study," Sports (Basel), vol. 11, no. 7, p. 131, Jul. 2023.

[12] K. Nakamura, M. Uchida, and T. Sato, "Basic research on the primary prevention of boxing-related sports injuries with the development of a quantitative motion analysis software," J. Phys. Ther. Sci., vol. 33, no. 6, pp. 495–498, Jun. 2021.

[13] D. Zhang, L. Bei, J. Wu, W. Li, and K. Zhang, "Effect of boxers' social support on mental fatigue: Chain mediating effects of coach leadership behaviors and psychological resilience," Work, 2023.

[14] R R. Kumar, "A comparative study of self-confidence among boxers and wrestlers of Hyderabad in India," *Int. J. Phys. Educ. Sports Health*, vol. 2, no. 1, pp. 1–4, 2015.

From Tables to 3D-Models: Improving the Usability of Scientific Data Visualization

Dennis Marschall*, Michael Butzek*, Nikolaos Margaritis*, and Ghaleb Natour*†
*Institute of Technology and Engineering (ITE), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany
†Welding and Joining Institute, Faculty of Mechanical Engineering
RWTH Aachen University, 52062 Aachen, Germany
email: [d.marschall, m.butzek, n.margaritis, g.natour]@fz-juelich.de

***Abstract*—In research facilities, critical components such as reactors or other centrepieces are equipped with a large number of sensors to record safety, process, and research-relevant data in real time. The effective visualisation of this measurement data is crucial for fast acquisition, analysis, and decision-making by plant operators. This Work in Progress Paper investigates different methods of visualising complex sensor data using the example of a 2-in-1 reactor, starting from classical tabular approaches to color-coded 2D-heatmaps, and interactive 3D-models. One focus is on human perception and the cognitive processing of the large amounts of data from 171 temperature measuring points. While tabular visualisations offer maximum precision, they are often unsuitable for the rapid detection of patterns and anomalies. Graphical visualisations - especially color-coded maps and 3D-models - provide an intuitive representation, but can be challenging due to color perception problems, information overload, and limited scalability. The system implementation, featuring an interactive 3D-model of the reactor, is described in detail in order to improve usability, reduce the cognitive load on the user and increase situational awareness. In further steps, specific usability tests will be carried out to validate the effectiveness of the 3D-visualisation and to analyze its influence on the user.**

Keywords-research facilities; HMI-Systems; 3D-integration.

I. Introduction

Research equipment often consists of a centerpiece, which is the primary focus of scientific investigations and serves to generate key measurement data. It is usually embedded within a complex environment of plant technologies, automation systems and a Human-Machine Interface (HMI). Such centerpieces can include detectors in accelerator systems, electric fields inside microscope probes, flow measurements in neutron targets [1], or temperature distributions in batteries, turbines, reactors, or other components.

The visualisation of this measurement data is crucial, particularly in research facilities, where fast comprehension and accurate interpretation are required. While traditional tabular visualisations provide precision, they often lack clarity when it comes to recognising spatial patterns or anomalies quickly. Alternative visualisation methods, such as 2D-heatmaps, and 3D-models, can support more intuitive data exploration. However, the choice of visualisation technique has a significant impact on usability and user performance.

In the field of Human-Computer Interaction (HCI) and visual analytics, previous work has investigated the cognitive effects of different visualisation forms. For instance Tory and Moller [2] emphasises that while 3D-representations may enhance spatial understanding, they can also introduce interaction complexity and perceptual challenges. Ware [3] highlights that, in many analytical tasks, well-designed 2D representations with multi-dimensional icons can be more effective than full 3D-renderings.

Despite these findings, 3D-visualisation can offer advantages in scenarios where data is inherently distributed across three dimensions, such as the temperature distribution in layered reactor systems. However, empirical validation of these approaches in research environments is still limited. This paper contributes to this gap by presenting the implementation of an interactive, browser-based 3D-visualisation for a 2-in-1 reactor.

The implementation addresses 171 temperature measurement points across three spatial planes, focusing on enhancing usability, reducing cognitive load, and supporting analytical tasks such as anomaly detection and system overview. Section II outlines the general visualisation requirements, followed by Section III, which compares different visualisation strategies. The implementation of the 3D-model is detailed in Section IV. Section V concludes with an outlook on planned usability evaluations and future applications in similar research systems.

II. Requirements

Different application scenarios also lead to different priorities in the requirements for the presentation of information. In applications where, for example, critical temperatures or temperature anomalies need to be recognised quickly, the quick and clear presentation of information for the user should take centre stage. At the same time, the system should offer interactive elements that allow the user to quickly show or hide details as required in order to focus attention on the relevant information. The ability to customise or filter the display according to individual requirements contributes to the flexibility of the system and ensures that the user has the most important data quickly available at all times [4].

For stable processes, on the other hand, the precise and accurate display of temperature values is paramount in order to make well-founded decisions. Here, a detailed, interactive display of temperature curves and the option of customised filtering or data analysis could support the user - for example,

by selecting specific time periods or parameters. The ability to customise the display of information to the respective requirements promotes efficient and effective use of the system [5].

The balance between precision and quick recognisability can change quickly depending on the situation, so scalability between overview and detail levels is usually required. The user should be able to switch between a quick, clear view and a detailed, precise display, depending on the type of information required at the time. In addition, all individual requirements should be easily adjustable without compromising user-friendliness. However, all individual requirements should follow basic ergonomic aspects of human-centred design in order to create a task-oriented HMI-System that allows the user to work effectively and efficiently [6]. Simple interactivity and customisable filter options help to reduce the workload and adapt the HMI-System flexibly to different needs and working conditions.

III. Realisation

The following section deals with the temperature recording and visualisation of a reactor, as it can be found, for example, in the field of hydrogen storage in research facilities. This consists of six radius levels on which a total of 86 internal heating tubes are arranged at the same distance from each other. For temperature measurement, 17 multi-thermocouples are guided onto the inner tubes from the outside, which measure the contact point of the inner tube, as well as other measuring points between the outer wall of the reactor and the inner tube, shown in Figure 1.

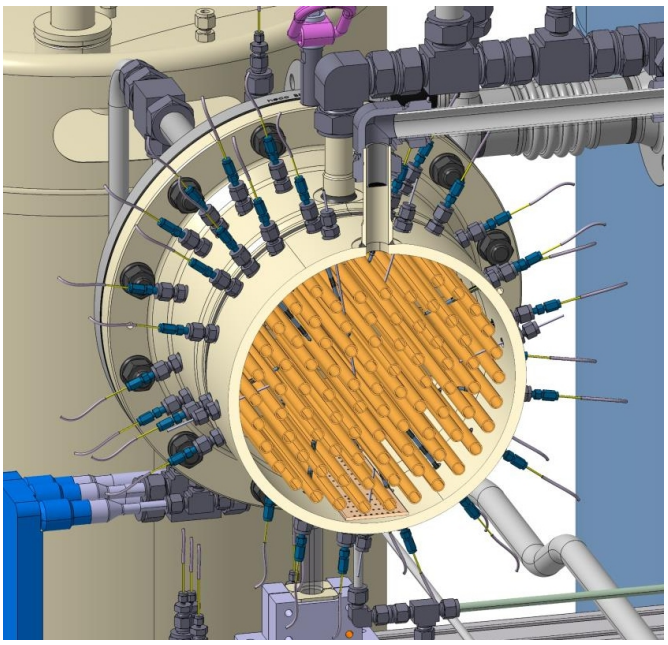

Fig. 1. Sectional view through the 2-in-1 reactor with a view onto the inner heating tubes and the multi-thermocouples.

This results in 17 measuring points that reflect the temperature of the inner tubes, and 40 measuring points of the temperature inside the reactor. This arrangement of thermocouples is placed three times along the length of the reactor, resulting in a total number of 171 temperature measuring points. The temperatures recorded in this way must be displayed to the system operator for various purposes, such as troubleshooting, monitoring against overheating and homogeneity. The various visualisation options are compared in the following subsections.

A. Tabular presentation

One of the most common methods of visualising temperature can be in tabular form. This enables a clear numerical representation of the temperature values. The advantages are the high accuracy and comparability of individual values, but it is no longer possible to assign them to the measuring position in the reactor without an additional drawing. The human ability to absorb information is limited, so tables with more than 5-9 entries (chunks) can cause difficulties [7]. Therefore, tables have the disadvantage that they are not very clear and it is difficult to recognise patterns or anomalies. Figure 2 shows the representation of the measured values of measuring plane-A from the component view of a thermocouple. The contact points on the heating tubes are shown in the left-hand column and the measuring points inside the reactor are shown in the following columns.

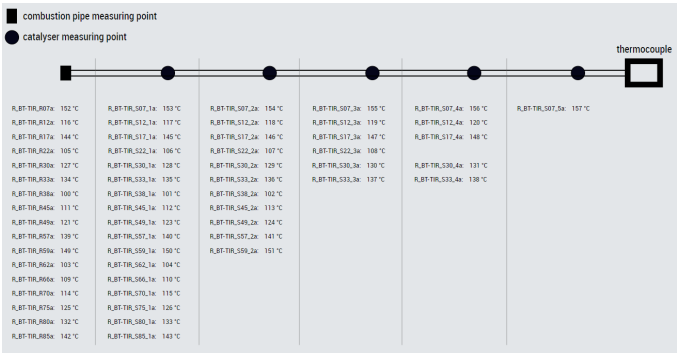

Fig. 2. Table based temperature monitoring of plane-A with 57 measuring values.

Despite the listed disadvantages of simplicity, this view can be very helpful in the application case of troubleshooting due to the direct assignment between the name of the measuring point and the current temperature value. A highlighting of limit value violations or other events can improve the interpretation of information.

B. 2D-presentation

If the table view of plane-A is transformed into a two-dimensional sectional view of the reactor, shown in Figure 3, the clarity and the possibility of recognising patterns or anomalies can be significantly improved.

In addition to the textual display of the temperature directly at the measuring point, the object can also be color-coded according to a defined color scale. The direct assignment of

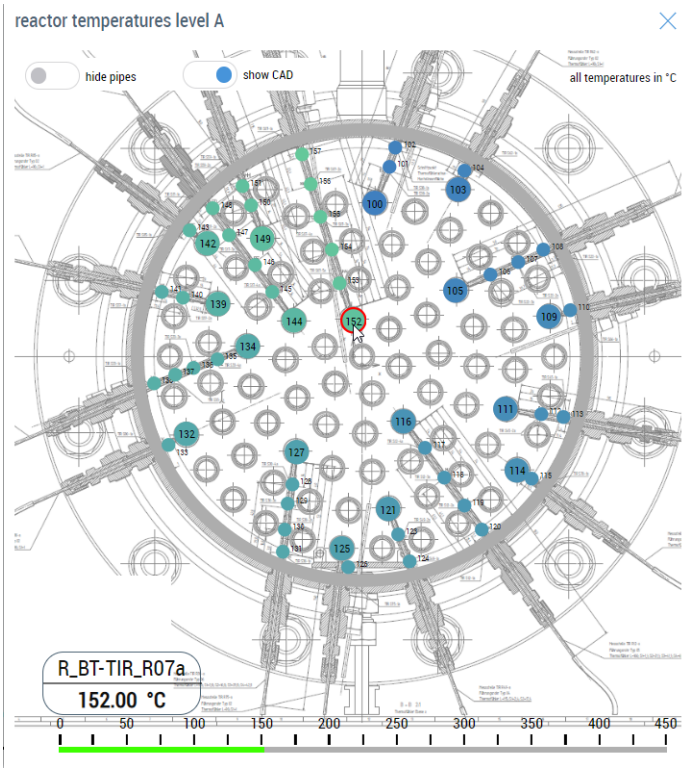

Fig. 3. 2D-based temperature monitoring of plane-A inside the construction drawing.

the temperature to the location inside the reactor, as well as the preattentive processing of colors, the ability of humans to recognise colors and color gradients very quickly [8], can be of great importance in the interpretation of local hotspots or cold areas. By adjusting the scale start and end values, the representation can be easily scaled to different temperature ranges. Nevertheless, color coding is only a quantitative classification of the measured value, as it is subject to errors due to different color perceptions and an indirect interpretation via the color scale. This disadvantage can be compensated by the additional textual representation, but the visual possibilities are much more limited by using the 2D-sectional view than in a tabular representation.

C. 3D-presentation

The two display types described refer to the measured values of one measuring plane. If the entire temperature curve within the reactor needs to be analysed, it is necessary to display all three planes simultaneously or to switch between the screens. This leads to a further loss of clarity due to the tripling of measuring points. The color-coded representation of temperatures in 2D-view can also be extended to the model in 3D-space. Figure 4 shows the simplified 3D-model of the reactor, in which the heating pipes are divided into three cylinder sections along their length and the measuring points within the reactor are represented by spherical objects.

The color of the object can then be converted into a temperature of the respective measuring point using the

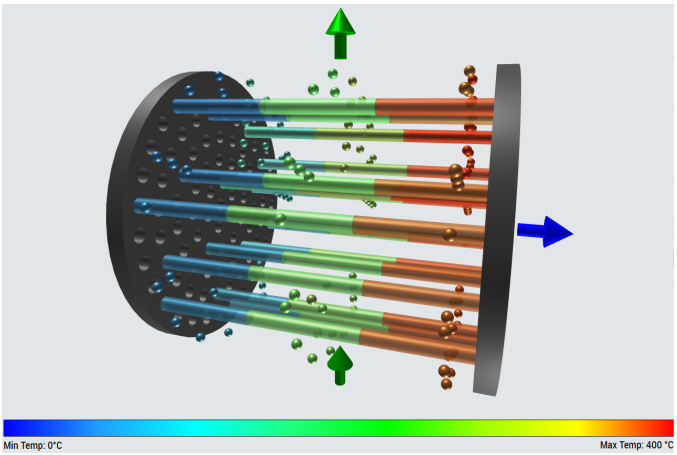

Fig. 4. 3D-based temperature monitoring of plane-A, B and C in the developed 3D-tool.

color scale, analogous to the 2D-representation. The 3D-display offers the option of moving the model, selecting different zoom levels and hiding individual objects in order to display the desired observation area. Clicking on individual objects inside the 3D-model opens a dialogue with the exact temperature of the object and additional properties. The 3D-visualisation of measured values, such as temperature profiles for example, offers an excellent overview, as it allows users to capture visual relationships intuitively. However, 3D-visualisation is not suitable for all types of measurement data or applications. In cases where the focus is on precise individual values, 2D-diagrams or tabular visualisations can often be more efficient. In addition, an overloaded 3D-visualisation can make it difficult to absorb information, especially when a large amount of data is displayed simultaneously.

IV. Implementation of 3D-integration

For the integration of 3D-content in the TwinCAT HMI, a 3D-Framework Controls was designed by Beckhoff Automation GmbH & Co. KG [9], which uses the JavaScript library Three.js [10] to display 3D-graphics inside the browser. The camera position, lights and colors, as well as scaling and movements can be influenced. Various functions are available, such as controlling animations or dynamically changing the colors, size or position of objects, which can be controlled from external Programmable Logic Controller (PLC) variables.

In order to implement a new model, it must be prepared in 3 steps. The model to be displayed was either designed directly in a Computer-Aided Design (CAD)-programme or is available as a .step file format. The model must first be broken down into individual assemblies, whereby each individual part to be edited later in the HMI must represent a separate assembly. These assemblies must then be exported to stereo lithography (STL) files, which simplifies the model into a surface model with triangular shapes [11]. Secondly, the individual .stl objects must be assembled into an overall model

in the 3D-tool Blender [12]. To do this, the coordinates from the original CAD-tool can be used, which are created during the separation into assemblies. All objects must be described with a unique name and objects for which the color is to be dynamically adjusted later need to be provided with a material property and a default color. The revised model must then be exported as .gltf format, a JSON-based 3D-format witch was chosen due to its good performance in web applications and its compatibility with the most commonly used browsers Firefox, Chrome and Microsoft Edge [13].

In the third and final step, a configuration file must be created and linked for the framework control, in which the 3D-model is embedded and all required dynamic functions are linked to PLC variables.

Figure 4 shows the representation of the processed reactor model from 177 assemblies in the HMI, in which the following functions were implemented:

- Display of the current temperature with a color-indicated temperature scale of 171 measuring points.
- Interaction with the model to open additional measuring point information.
- Whitening of passive components through opacity.
- Insert visual section planes to display the 3 planes of the measuring points.
- Opening defined views and scalings.
- Activating the auto-rotate function and influencing the rotation speed.

V. Conclusion and future work

The visualisation of measurement data in research facilities poses unique challenges: it requires both the rapid identification of critical states and the precise interpretation of complex data. This paper presents a modular 3D-visualisation approach integrated into an HMI for a 2-in-1 reactor, based on the Three.js framework. The implementation demonstrates the feasibility of mapping 171 temperature sensors onto a spatially accurate 3D-model, providing color-coded data overlays, interactive views, and real-time control via PLC variables.

While the current solution offers an intuitive and immersive experience, its effectiveness for specific analytical tasks remains to be evaluated. Previous research in visual analytics and HCI suggest that while 3D-representations support spatial understanding, they can be less effective than 2D methods for certain analytical tasks [3]. In future work, usability tests will be conducted with domain experts to assess the cognitive load, task efficiency, and user preference between tabular, 2D, and 3D-visualisations. These tests will also consider accessibility aspects, such as color vision deficiencies, which are not yet fully addressed in the current implementation.

Furthermore, the system will be extended to support other types of sensor data and centerpieces. An important part of future work is to derive task-specific guidelines for when 3D-visualisation is beneficial and when simplified, abstracted views may be more effective.

References

[1] Y. Beßler, "Fluid mechanical simulation and experimental validation of the cryogenic hydrogen-moderator for the european spallation neutron source ESS," Fakultät für Maschinenwesen der Rheinisch-Westfälischen Technischen Hochschule Aachen, 2020.

[2] M. Tory and T. Moller, "Human factors in visualization research," in IEEE Transactions on Visualization and Computer Graphics, vol. 10, no. 1, pp. 72-84, Jan.-Feb. 2004. https://doi: 10.1109/TVCG.2004.1260759.

[3] C. Ware, "Information Visualization: Perception for Design: Second Edition," Morgan Kaufmann Publishers, 2004.

[4] D. Norman, "The Design of Everyday Things," Basic Books, 2013.

[5] B. Shneiderman et al., "Designing the User Interface: Strategies for Effective Human-Computer Interaction," Pearson, 2017.

[6] DIN EN ISO 9241-210, "Ergonomie der Mensch-System-Interaktion – Teil 210: Menschzentrierte Gestaltung interaktiver Systeme," "Ergonomics of human-system interaction - Part 210: Human-centred design of interactive systems," (ISO 9241-210:2019).

[7] G.A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," Psychological Review, 63(2), pp. 81–97. 1956.

[8] C. Healey and J. Enns, "Attention and Visual Memory in Visualization and Computer Graphics," in IEEE Transactions on Visualization and Computer Graphics, vol. 18, no. 7, pp. 1170-1188, 2012.07.

[9] Beckhoff Automation GmbH and Co. KG, "Dokumentation zur Konfiguration des Framework Controls zur Anzeige eines 3D-Modells," "Documentation for configuring the framework control to display a 3D model," 2023.04.

[10] B. Danchilla, "Three.js Framework," In: Beginning WebGL for HTML5, Apress, Berkeley, CA, 2012.

[11] S.H. Huang, L.C. Zhang, and M. Han, "An Effective Error-Tolerance Slicing Algorithm for STL Files," Int J Adv Manuf Technol 20, pp. 363–367 (2002). https://doi.org/10.1007/s001700200164.

[12] "Blender 4.3 Reference Manual," https://docs.blender.org/manual/en [Online], Accessed: 2025.02.07 Last updated 2025.01.10.

[13] G. Lee et al., "A Study on the Performance Comparison of 3D File Formats on the Web," International journal of advanced smart convergence, vol. 8, no. 1, pp. 65–74, 2019.03.

Improving Continuous Japanese Fingerspelling Recognition with Transformers: A Comparative Study against CNN-LSTM Hybrids

Akihisa Shitara†*, Yuhki Shiraishi*
†Graduate School of Library, Information, and Media Studies, University of Tsukuba, Japan
Email: theta-akihisa@digitalnature.slis.tsukuba.ac.jp
*Faculty of Industrial Technology, Tsukuba University of Technology, Japan
Email: yuhkis@a.tsukuba-tech.ac.jp

***Abstract*—To achieve smooth communication between d/Deaf and hard of hearing (d/DHH) and hearing people, we have developed a continuous Japanese Fingerspelling (JF) recognition system using sensor gloves and deep learning. We have selected a light and inexpensive sensor glove adapted for the system's daily use. In our prior system using a machine learning model that combines Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM), despite achieving an average micro F-measure of 92.1% across the 76 JF characters, we reported the average macro F-measure of only 64.7%. Two problems cause this issue: distinguishing between static and dynamic fingerspellings, and the decreased recognition rate due to the large number of instances "ϕ" (the transition movements characters). Therefore, we conducted a quantitative evaluation using the CNN-LSTM combined machine learning model as a baseline to verify whether the Transformer Encoder could improve JF recognition rates. Consequently, for the 76 JF characters, the average micro and macro F-measures were 93.8% (0.2) and 77.4% (1.0), respectively.**

Keywords–Deaf and hard of hearing; Sign language; Sensor glove; Recognition.

I. INTRODUCTION

To achieve smooth communication between d/Deaf and Hard of Hearing (d/DHH) and hearing people, we have developed a continuous Japanese Fingerspelling (JF) recognition system using sensor gloves and deep learning [1] [2]. However, a machine learning model that combines Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) [2] despite achieving the average micro F-measure of 76 JF characters was 92.1%, we reported the average macro F-measure of only 64.7% (Figure 1). Two problems cause this issue: distinguishing between static and dynamic fingerspellings, and the decreased recognition rate due to the large number of instances "ϕ" (the transition movements characters). However, in the current research community on sign language recognition, a machine learning model based Transformer [3] is often used. Thus, we conduct a comparative analysis by a quantitative evaluation using the CNN-LSTM combined machine learning model as a baseline to verify whether the Transformer Encoder could improve JF recognition rates. Additionally, in the current research community on fingerspelling recognition and sign language recognition, the small size of the dataset is listed as an issue. To the best of our knowledge, no reports have verified whether there are individual differences among signers expressing continuous fingerspellings. Thus, we also compare and analyze the impact of individual differences among signer data by conducting cross-validation evaluations selecting training data.

The remainder of this study is organized as follows. In Section II, we describe related studies on fingerspelling recognition and sign language recognition. In Section III, we describe a comparative analysis by a quantitative evaluation using the CNN-LSTM combined machine learning model as a baseline. In Section IV, we describe a comparative analysis learning model as CNN-Transformer Encoder as a baseline. In Section V, we discuss the results and the limitations of the work. Finally, in Section VI, we provide some concluding remarks and suggest some avenues for future research.

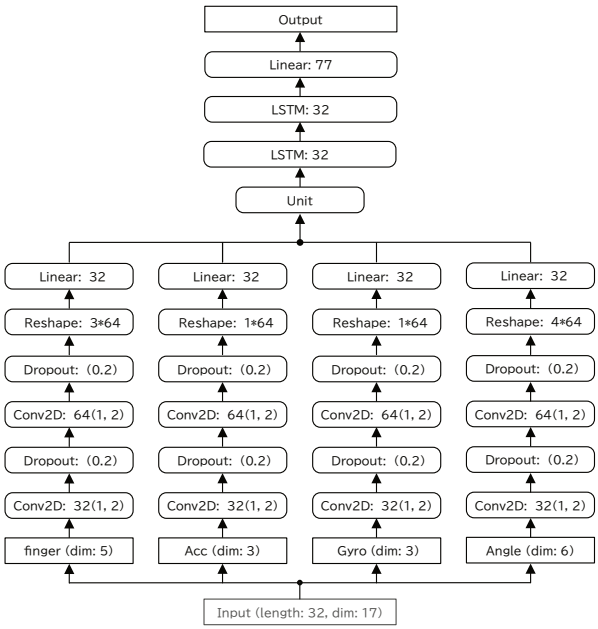

Figure 1. Architecture of the CNN-LSTM combined machine learning model.

II. RELATED WORK

Previous research on fingerspelling and sign language recognition has proposed two types of sensors for recognizing a series of operations in fingerspelling and sign language: contact-type sensor gloves and non-contact-type cameras.

A. Image recognition

Several methods have been proposed for recognizing hand shapes based on processing images of fingerspelling as captured by cameras.

As example of a fingerspelling recognition method, Mukai et al.'s method [4] used a classification tree and machine learning based on a support vector machine to classify individual

images; it targeted 41 immobile characters in Japanese Sign Language (JSL) resulted in an average recognition accuracy of 86%. Hosoe et al.'s method [5] used deep learning for recognition and achieved a recognition rate of 93%, but only for static fingerspellings. Jalal et al.'s method [6] achieved a recognition rate of 99% for American Sign Language (ASL) images based on a deep learning algorithm for static fingerspellings (i.e., excluding "J" and "Z").

However, the recognition accuracy could not be considered as sufficient for practical recognition in JF. Additionally, relatively few recognition results have been reported for dynamic fingerspellings (i.e., fingers moving when expressing a character). For example, in Kondo et al.'s study [7] of dynamic fingerspellings in JSL, the identification of hand shapes was performed using a kernel orthogonal mutual subspace method from images of hand regions obtained from distance images, and the classification of movements was performed using decision trees based on center-of-gravity coordinates. These results yielded a 93.8% identification rate. However, the recognition accuracy was insufficient for the practical recognition required for JF.

Furthermore, examples of machine learning models based on Transformers for sign language recognition include the machine learning model SignAttention [8] targeted Greek Sign Language, the machine learning model [9] using the American Sign Language dataset How2Sign [10], the machine learning model [11] using the German Sign Language dataset RWTH-PHOENIX-Weather [12] [13] and other examples such as the machine learning model [11]. Moreover, as a result survey a related research on sign language recognition [14], it is reported that CNN, LSTM, and Transformer are used in many research, and as reported in the research that surveyed the State-Of-The-Art (SOTA) in sign language recognition [15], it is reported that Transformer is used in many cases.

On the other hand, there is also research [16] using combination spatial-temporal modules and Multi-Layer Perceptron (MLP), and Takayama et al.'s model [17] using combination Spatial Temporal Graph Convolutional Networks [18] and Transformer and targeted JSL. Additionally, as examples use Conformer [19], Kimura et al.'s machine learning model [20] targeted JSL. Signformer [21] also used a module that has been redesigned based on the Conformer architecture.

B. Sensor glove recognition

Several methods have been proposed for recognizing hand shapes based on measurement data acquired by contact-type sensor gloves. These methods can measure data, which includes the flexion of the five fingers, the position and direction of the hand. The measurement data are then sent to a personal computer or microcomputer, and a classification algorithm is used to recognize hand shapes.

As example of a fingerspelling recognition method, Cabrera et al. [22] used the Data Glove 5 Ultra [23] sensor glove with an acceleration sensor to acquire information regarding the degree of flexion of each finger and wrist direction. The study targeted 24 static fingerspelling characters in ASL, excluding "J" and "Z", and achieved a recognition rate of 94.07%. Mummadi et al. [24] prototyped a sensor glove with multiple embedded inertial sensors. The study targeted fingerspellings of French Sign Language, and achieved an average recognition rate of 92% with an F1-score of 91%. Kakoty et al.'s model [25] used kernel-supported vector machine, and targeted a dataset of one-handed Indian sign language alphabets (C, I, J, L, O, U, Y, W), ASL alphabets (A to Z), and signed numbers (0 to 9). The study achieved an average recognition rate of 96.7%. SpellRing [26] used combined active acoustic sensing and Inertial Measurement Unit (IMU) in the ring-shaped devices worn on the thumb. ResNet-18 [27] uses CNN as the backbone and also leverages Connectionist Temporal Classification (CTC) [28].

Furthermore, as examples of a sign language recognition method, Chong et al. [29] placed six IMUs on the back of the palm and on each fingertip to capture their motion and orientations. The study targeted 28 proposed word-based sentences in ASL, and used LSTM. The method achieved an accuracy of up to 99.89%. SmartASL [30] uses IMUs installed in two devices, an earphone and a smartwatch, to include not only manual marker expressions but also Non-manual Marker (NM) expressions. The method used LSTM for data related to hand movements and CNN and LSTM for data related to NM expressions. After that, the Transformer model T5 [31] is used for fine-tuning together with translated English sentences. SignRing [32] uses the IMUs in the ring-shaped devices worn on the index fingers of both hands. It generates data similar to that from the IMU sensors from the sign language videos in the ASLLRP [33] ASL dataset. Then, it uses a model combining CNN and LSTM to train the generated data.

III. First Comparative Analysis

To verify whether the JF recognition rate improved when replacing LSTM with a Transformer Encoder, and to compare the impact of the differences between machine learning model architectures, we trained each model and compared their accuracies. We have the following machine learning models:

- 2LSTM
- branch-2CNN-unit-2LSTM
- Transformer Encoder
- branch-CNN-unit-Transformer Encoder
- branch-2CNN-unit-Transformer Encoder

A. Continuous Japanese fingerspelling dataset

The continuous Japanese fingerspelling dataset used in this study is our previously collected data [2]. The sensor glove used for our previous data collection consists of Arduino Pro Mini and MPU6050, where conductive fiber weaving techniques [34] detect finger movements based on resistance changes, and the MPU6050 detects acceleration and angular velocity. The dataset contains words consisting of three to five fingerspelling characters, with each word comprising 11 dimensions (finger movement: five dimensions, acceleration: three dimensions, angular velocity: three dimensions) $\times$ 960 samples (120 sps $\times$ 8 sec). The dataset includes data from 33 participants, with each person performing 64 words five times each. As preprocessing, we calculated angles using the angular velocity three dimensions and added angles six dimensions ($\sin$ and $\cos$). Using moving average calculations, we also reduced the 960 samples (120 sps $\times$ 8 sec) to 32 samples (4 sps $\times$ 8 sec). Therefore, the input dimensions are 32 samples $\times$ 17 dimensions (length: 32, dim: 17).

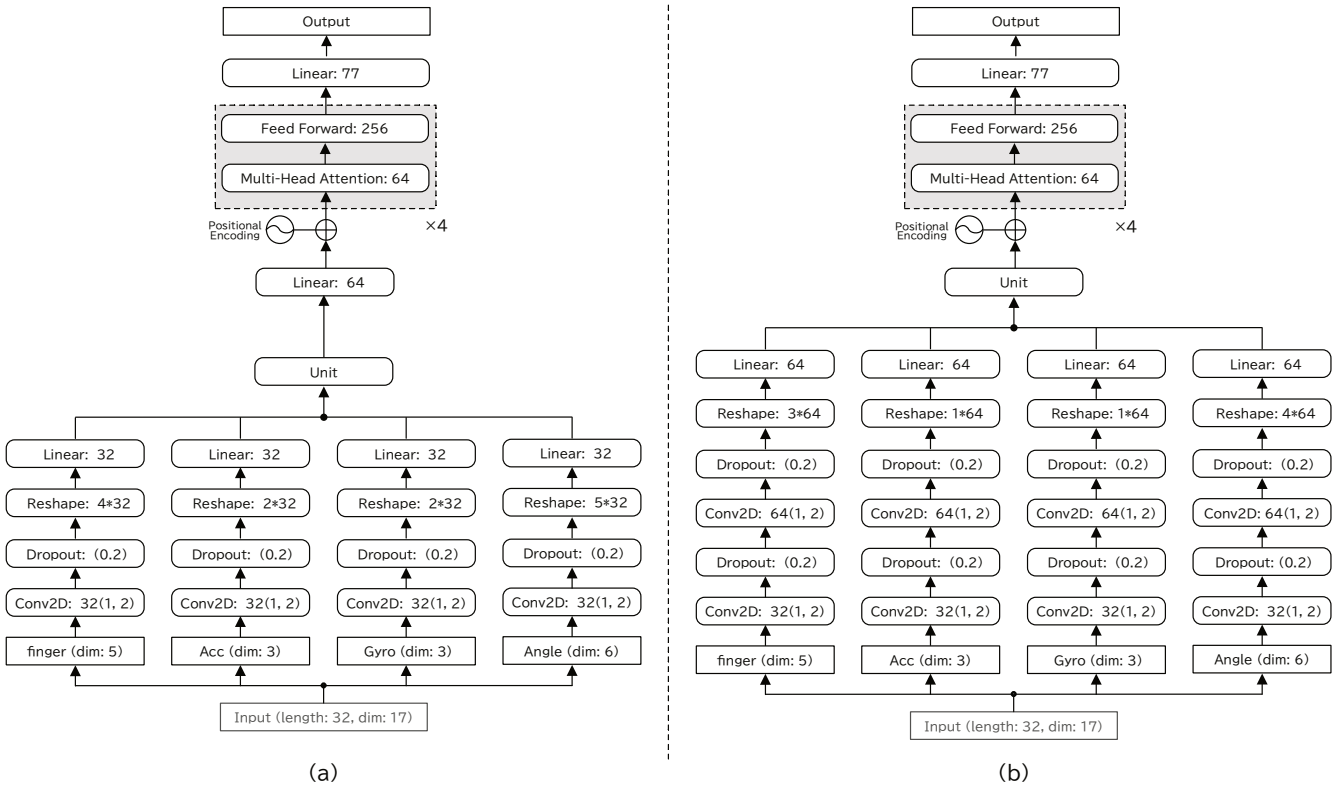

Figure 2. (a) Machine learning model architecture of the "branch-CNN-unit-Transformer Encoder", (b) Machine learning model architecture of the "branch-2CNN-unit-Transformer Encoder".

B. Machine Learning Model Architecture

For the development environment, we used a Docker container image distributed by NVIDIA [35]. The main specifications are as follows. We rebuilt all models, including our previously constructed branch-2CNN-unit-2LSTM, switching from TensorFlow to PyTorch.

- Ubuntu 22.04
- NVIDIA CUDA 12.3.0
- Python 3.10
- PyTorch 2.2.0a0+6a974bec

This motive is verify whether the JF recognition rate of both 2LSTM and Transformer Encoder when combined with CNN, improved compared to when not combined with CNN. Therefore, we included both 2LSTM, which consist of two LSTM layers, and the Transformer Encoder in this comparative analysis. Next, Figure 2 (a) and (b) show the machine learning model architectures that combine one and two layers of CNN with the Transformer Encoder. The reason for this architecture is that, similar to branch-2CNN-unit-2LSTM, the data is split into separate branches for each feature dimension, such as the finger movement, acceleration, angular velocity, and angle, and then input through the CNN layer.

C. Evaluation experiments

We evaluated each machine learning model architecture (Table I). First, we set the epoch to 3,000, and set the patience to 300 for stopping training using EarlyStopping as a measure against overfitting. In addition, we set the batch size to 64 for 2LSTM and branch-2CNN-unit-2LSTM, 16 for Transformer Encoder, and 32 for branch-CNN-unit-Transformer Encoder and branch-2CNN-unit-Transformer Encoder because the batch size was set to the best accuracy in each machine learning model from the result we tested at the batch size 16, 32, and 64, respectively.

1) Comparison of k-Fold Cross-Validation Methods: We evaluated each model using the 5-fold Cross-Validation (CV) and the 10-fold CV. The input data was shuffled and then divided into training and test data, and Table I shows the average and standard deviation of the results of the 5 and 10 runs for each model. Moreover, Table I values are not the values at the epoch when learning was stopped due to early stopping to prevent overfitting, but the values at the epoch when the validation loss was minimized. As shown in Table I, the F-measure micro of each model at the 5-CV and the 10-CV, except for 2LSTM, is over 90%. Comparing Transformer Encoder to 2LSTM, the F-measure micro and macro improved over that of the former. The same improvement was observed in comparing Transformer Encoder combined with CNN to 2LSTM combined with CNN. In particular, the F-measure macro for Transformer Encoder, branch-CNN-unit-Transformer Encoder, and branch-2CNN-unit-Transformer Encoder is improved by nearly 10% when compared to 2CNN-2LSTM, suggesting that the decreased recognition rate due to the large number of instances "ϕ" (the transition movements characters), which was a previous study [2] issue, has been alleviated.

TABLE I. MODEL COMPARISON EVALUATIONS IN FIRST COMPARATIVE ANALYSIS. THE PARENTHESES INDICATE THE STANDARD DEVIATION.

machine learning model architecture	k=5, F-measure [%]		k=10, F-measure [%]		k=33 (*random*[a]), F-measure [%]		k=33 (*person*[b]), F-measure [%]	
	micro	macro	micro	macro	micro	macro	micro	macro
2LSTM	89.9 (0.2)	50.5 (1.2)	90.2 (0.3)	51.6 (1.6)	90.4 (0.4)	52.1 (2.1)	88.8 (2.2)	40.5 (9.6)
branch - 2CNN - unit - 2LSTM	91.9 (0.1)	64.6 (0.7)	92.1 (0.2)	65.4 (1.1)	92.2 (0.4)	66.0 (2.5)	90.0 (0.3)	50.7 (13.1)
Transformer Encoder	92.8 (0.3)	72.5 (1.3)	93.3 (0.3)	74.9 (1.8)	93.5 (0.5)	75.8 (2.4)	92.2 (2.8)	69.1 (10.3)
branch - CNN - unit - Transformer Encoder	93.4 (0.1)	75.8 (0.6)	93.6 (0.2)	76.5 (0.8)	93.8 (0.5)	77.7 (2.2)	92.4 (2.9)	70.8 (10.1)
branch - 2CNN - unit - Transformer Encoder	93.3 (0.2)	74.8 (0.9)	93.6 (0.3)	76.6 (1.0)	93.8 (0.4)	77.1 (2.2)	92.4 (2.7)	70.8 (9.1)

[a] *random*: evaluation using randomly selected data for 33-fold CV.
[b] *person*: evaluation where the evaluation set consists of data from a single individual.

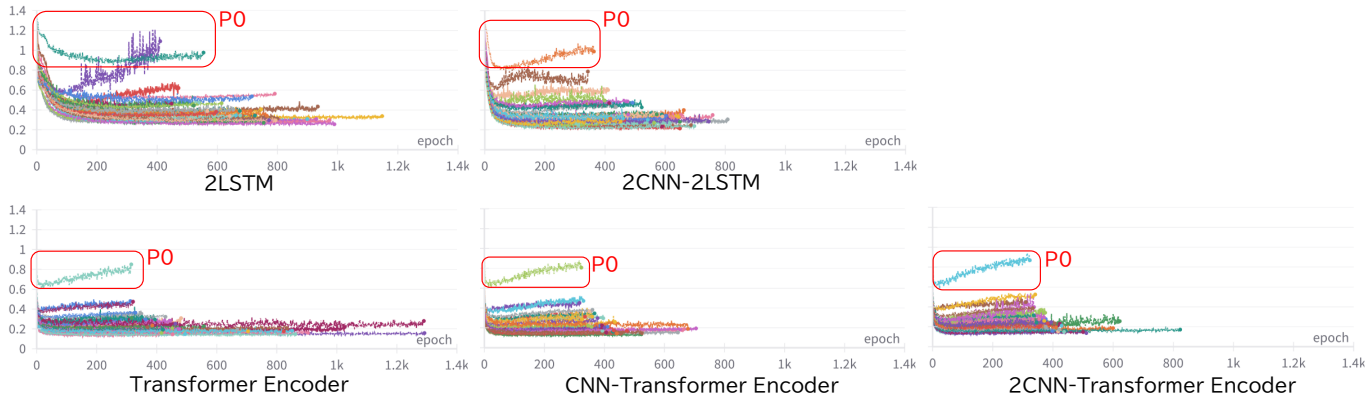

Figure 3. Learning progress graph for each model (horizontal axis: number of epochs, vertical axis: validation loss): The loss for each person in the 33-fold CV method is shown as a line graph, and it was observed that the data for one person (P0) showed a significant deviation from the others in the vicinity of 0.6 and 0.8.

2) The Impact of Individual Differences on 33-Fold Cross-Validation: We evaluated the results using the 33-fold CV, first, with case the input data for one person used as test data and the data for the remaining 32 people used as training data (k=33(person) in Table I), and second, the input data was shuffled and then divided into training and test data (k=33(random) in Table I). Moreover, Table I values are not the values at the epoch when learning was stopped due to early stopping to prevent overfitting, but the values at the epoch when the validation loss was minimized. As described in Section III-C1, the macro average of the F-measurement improved for all three models (Transformer Encoder, branch-CNN-unit-Transformer Encoder, and branch-2CNN-unit-Transformer Encoder) compared to branch-2CNN-unit-2LSTM. However, we found that we needed to consider the significant validation loss for the same person's test data. Figure 3 shows the graph showing the change in validation loss for each model at the 33-fold CV with case the input data for one person used as test data and the data for the remaining 32 people used as training data.

Table II presents a comparative analysis of F1-scores for individual fingerspelling characters across three models: the previous 2CNN-2LSTM model and the two best-performing models (CNN-Transformer Encoder and 2CNN-Transformer Encoder), under three conditions: random data distribution, P0, and P1. A notable improvement was observed in the recognition of challenging characters. While the 2CNN-2LSTM model showed zero F1-scores for 28 characters in the P0 condition, this number significantly decreased to 5 and 4 characters for the CNN-Transformer Encoder and 2CNN-Transformer Encoder models, respectively. Furthermore, the 2CNN-Transformer Encoder under P1 condition demonstrated robust performance, with no characters receiving zero F1-scores, outperforming both the 2CNN-2LSTM and CNN-Transformer Encoder models.

IV. Second Comparative Analysis

Using branch-CNN-Transformer Encoder and branch-2CNN-Transformer Encoder that showed good accuracy in First Comparative Analysis, we examine the impact of the input data for participant P0 identified in the impact of individual differences validation comparison. We also examine the impact of adding BatchNorm-1D and Rectified Linear Unit (ReLU) to each of the two machine learning model architectures. The timing of adding BatchNorm-1D and ReLU is when inputting to the Transformer Encoder module from the CNN module.

A. Evaluation experiments

For each branch-CNN-Transformer Encoder and branch-2CNN-Transformer Encoder, we evaluated the results using the 10-fold CV with three different combinations: add BatchNorm-1D and ReLU, removing the input data for participant P0, applying both modifications. The input data was shuffled and then divided into training and test data, and Table III shows the average and standard deviation of the results of the 10 runs for each model. Moreover, Table III values are not the values at the epoch when learning was stopped due to early stopping to prevent overfitting, but the values at the epoch when the validation loss was minimized. We also included the 10-fold CV values from Section III-C1 to check improvement from the no-applying case. CNN-Transformer Encoder and 2CNN-Transformer Encoder showed improvement in the F-measure macro compared to the no-applying case from case removing

TABLE II. THE F1-SCORES FOR EACH OF THE FINGER CHARACTERS IN THE CHARACTERS IN THE MODEL COMPARISON IN THE FIRST COMPARATIVE ANALYSIS.

2CNN-2LSTM						CNN-Transformer Encoder						2CNN-Transformer Encoder					
k=33		P0		P1		k=33		P0		P1		k=33		P0		P1	
chi	41.3	ho	0.0	me	0.0	du	55.3	nu	0.0	nu	0.0	du	53.9	na	0.0	di	33.3
ho	44.2	bo	0.0	hu	0.0	di	59.7	da	0.0	yo	0.0	di	58.6	so	0.0	nu	50.0
pe	46.6	pa	0.0	du	0.0	chi	63.9	di	0.0	du	40.0	chi	63.8	ke	0.0	re	52.6
du	48.9	hu	0.0	re	21.1	nu	65.2	ze	0.0	yu	50.0	pe	64.3	di	0.0	ma	53.8
te	49.9	zi	0.0	xya	25.0	pe	67.1	ke	0.0	re	53.3	ho	68.6	pu	14.3	ho	54.5
pu	53.4	ha	0.0	ni	27.3	ho	68.0	ta	11.8	di	54.5	bo	70.0	ho	16.7	te	54.5
hu	53.7	no	0.0	ne	28.6	pi	69.9	pi	13.3	ra	54.5	so	70.7	ko	20.0	chi	54.5
he	54.5	ze	0.0	pe	28.6	tsu	71.4	ko	15.4	chi	54.5	he	70.8	a	22.2	xtsu	59.5
di	54.6	ni	0.0	di	37.5	yu	71.7	ha	17.4	ta	57.1	tsu	70.9	ze	23.5	ra	60.0
so	54.6	ro	0.0	yu	40.0	na	71.9	du	17.4	xyo	60.0	nu	71.8	chi	24.4	du	62.5
ni	55.9	pi	0.0	ho	40.0	ta	72.6	te	18.2	me	63.2	pi	71.9	du	26.1	yo	66.7
ko	56.0	te	0.0	he	44.4	se	72.7	e	19.0	ma	64.0	se	72.4	pi	26.7	yu	66.7
pi	56.3	xtsu	0.0	te	44.4	ro	72.8	wo	21.7	ne	66.7	yu	72.6	da	26.7	hu	66.7
ha	56.4	chi	0.0	wo	46.2	yo	72.9	o	25.0	de	66.7	na	73.0	pe	28.6	ni	66.7
ga	57.4	so	0.0	ro	47.6	he	73.1	ho	26.7	hu	66.7	ha	73.1	e	28.6	ta	66.7
ro	57.6	yu	0.0	nu	50.0	so	73.1	xya	26.7	ho	66.7	ta	73.2	ge	28.6	zu	66.7
ka	57.8	bi	0.0	chi	51.6	ha	73.2	ro	28.6	zo	66.7	o	73.9	wo	28.6	ka	66.7
pa	57.9	bu	0.0	xtsu	52.2	me	73.9	bu	28.6	ha	66.7	ro	74.3	ha	28.6	go	69.2
tsu	58.3	da	0.0	bo	53.3	de	73.9	so	28.6	pa	66.7	yo	74.4	tsu	30.5	xyu	69.6
bo	58.8	ko	0.0	ko	53.3	ko	74.0	pe	28.6	bo	70.6	hu	74.5	bu	30.8	ha	70.0
ne	60.3	ke	0.0	de	54.5	bo	74.0	go	30.0	pe	70.6	te	74.6	te	30.8	sa	70.2
xtsu	60.8	ga	0.0	ha	55.6	hu	74.0	tsu	30.3	hi	71.0	wo	74.9	me	31.3	xya	70.6
me	61.3	pu	0.0	pa	57.1	su	74.4	wa	31.6	ro	71.4	hi	75.0	ro	31.6	pe	70.6
zi	62.5	a	0.0	hi	57.1	xya	75.1	bi	31.6	ge	72.7	me	75.0	gu	33.3	zi	70.6
ya	62.5	di	0.0	ze	57.1	te	75.4	yo	31.6	tsu	72.7	re	75.3	o	33.3	ya	70.8
wa	62.5	wo	0.0	ra	57.1	ni	75.4	no	33.3	ka	72.7	xya	75.4	pa	33.3	de	71.4
nu	62.5	pe	0.0	ke	57.1	re	75.5	de	35.3	wa	73.7	ni	75.5	bi	33.3	pa	71.4
hi	62.6	du	0.0	xyu	58.3	da	75.5	ra	35.3	go	74.1	ge	75.5	yo	33.3	za	71.4
de	63.0	su	9.7	pu	58.8	ba	75.6	a	36.4	za	75.0	su	75.6	mo	34.4	ku	71.7
a	63.2	ru	9.7	tsu	58.8	hi	75.6	pu	37.5	wo	75.0	ne	76.2	yu	37.5	a	72.7
da	63.2	ta	10.0	go	61.5	ge	75.6	ge	37.5	xtsu	75.0	de	76.2	zo	37.5	gu	73.7
o	63.9	wa	13.8	sa	63.2	ke	75.8	yu	37.5	e	75.0	ba	76.3	ne	38.1	me	73.7
su	65.7	ne	14.3	po	64.7	o	76.5	pa	38.1	xya	75.0	bi	76.6	bo	40.0	wa	73.7
yu	66.0	de	14.3	be	66.7	wo	76.6	ne	38.1	ke	75.0	ma	76.9	nu	40.0	ne	73.7
ba	66.0	ri	14.6	da	66.7	ne	77.1	me	38.7	xyu	75.0	pa	76.9	ba	40.0	tsu	75.0
na	66.1	ra	18.2	yo	66.7	gu	77.4	zi	40.0	ru	75.3	xyu	77.1	ki	40.4	e	75.0
ma	66.2	xya	20.0	ma	66.7	ra	77.7	be	40.0	do	75.5	gu	77.7	zu	42.1	bo	76.2
shi	66.3	hi	21.4	ka	66.7	e	77.8	se	40.0	sa	75.5	ko	78.0	re	42.4	gi	76.7
bi	66.3	me	22.2	ta	66.7	ma	77.9	gu	41.0	su	75.9	da	78.0	xtsu	42.6	su	77.2
yo	66.7	ma	22.7	se	66.7	wa	77.9	su	41.6	ku	76.4	xtsu	78.4	su	43.3	no	77.2
po	66.9	ya	24.0	mo	69.1	bi	78.1	re	42.9	ya	76.6	to	78.6	ku	43.8	da	77.8
wo	67.2	go	25.5	za	69.2	ru	78.1	chi	43.5	mo	76.9	e	78.7	xya	44.4	ko	77.8
se	67.2	ba	26.1	wa	70.0	xtsu	78.2	ri	45.9	ni	78.6	ya	78.7	se	46.2	he	78.3
ta	67.3	tsu	26.8	zu	70.6	mo	78.2	gi	46.7	u	79.7	ru	78.8	hu	46.7	hi	78.8
sa	68.1	o	27.6	ya	71.1	ki	78.6	shi	47.1	te	80.0	ra	78.8	go	46.8	-	79.0
e	68.3	ge	28.6	zo	71.4	ya	78.7	ku	47.2	po	80.0	po	78.9	ra	47.1	wo	80.0
re	68.4	zo	29.6	bi	71.4	xyu	79.1	bo	47.6	so	80.0	shi	78.9	no	48.0	xyo	80.0
ru	68.5	re	30.0	ga	71.4	po	79.2	sa	47.9	a	80.0	u	79.0	ga	48.5	pu	80.0
xya	69.0	za	30.0	su	71.7	shi	79.3	hu	48.0	i	80.9	ka	79.1	ta	50.0	mo	80.0
be	69.1	na	30.8	ri	73.5	to	79.7	xtsu	49.1	to	81.0	mo	79.1	po	50.0	ke	80.0
ra	69.4	be	30.8	i	75.2	pa	79.9	ma	50.0	ri	81.3	bu	79.3	ni	50.0	zo	80.0
go	70.4	se	31.3	ru	75.6	a	79.9	ba	50.0	-	81.4	go	79.5	he	50.0	do	80.6
ze	70.8	gi	32.7	u	76.0	u	80.0	na	50.0	gu	82.4	a	79.6	hi	50.0	u	81.0
ge	70.9	he	33.3	to	76.1	ka	80.0	zo	50.0	o	82.4	ke	79.7	ri	50.3	se	81.3
ke	70.9	ku	35.0	e	76.2	i	80.4	mo	50.7	no	83.3	ki	80.2	za	51.9	po	81.5
gu	71.1	yo	35.3	mu	76.5	sa	80.6	ni	51.1	gi	83.6	wa	80.4	de	52.6	ru	82.5
zu	71.9	po	35.3	so	76.9	ga	80.6	ki	51.7	ko	84.2	sa	80.7	shi	52.6	to	83.2
bu	73.4	nu	35.3	ge	76.9	zu	80.6	ru	53.6	ze	85.7	zu	81.0	-	53.6	bu	83.3
i	73.7	do	37.5	zi	76.9	no	81.0	za	53.8	zi	85.7	i	81.0	u	54.2	bi	83.3
zo	73.8	shi	40.5	o	78.3	xyo	81.1	he	53.8	ga	85.7	do	81.5	zi	55.6	ro	83.3
mo	74.0	sa	40.6	do	79.3	go	81.2	-	54.1	ba	85.7	mu	81.5	i	56.0	shi	84.0
u	74.6	ka	40.7	ku	80.0	mu	81.2	zu	54.5	se	86.7	ga	81.6	nn	56.3	so	84.2
ki	75.3	-	41.4	-	81.0	bu	81.5	hi	55.0	he	87.0	no	81.6	ma	56.5	ri	85.4
xyu	75.8	u	43.9	gu	82.4	do	81.6	u	55.4	da	87.5	zo	81.7	gi	57.1	ba	85.7
no	76.1	e	44.4	shi	82.4	za	81.8	i	58.4	zu	87.5	gi	81.8	wa	57.1	i	85.7
za	77.1	i	45.5	gi	82.6	gi	82.0	nn	58.4	shi	88.0	ri	82.6	sa	57.5	nn	85.9
to	77.1	ki	47.9	no	83.0	pu	82.1	mi	58.9	nn	88.3	pu	82.6	ka	57.8	ze	87.5
gi	77.3	mo	48.5	xyo	83.3	ku	82.2	ka	61.2	pi	90.9	ze	82.9	to	58.8	ga	87.5
do	77.4	zu	50.0	nn	84.3	ri	82.8	xyu	61.5	ki	92.0	za	82.9	mi	60.4	mi	88.3
xyo	77.8	to	58.9	a	85.7	-	82.9	po	63.6	mu	92.3	xyo	83.0	ru	63.8	o	90.9
-	78.6	nn	59.2	ba	85.7	mi	83.2	xyo	64.0	be	94.1	zi	83.2	mu	65.2	ki	91.3
ri	78.7	xyu	59.5	ki	88.4	zo	83.3	to	64.8	pu	94.1	ku	83.4	be	66.7	mu	92.3
ku	78.9	gu	63.6	bu	88.9	nn	83.5	mu	65.4	mi	94.6	-	83.5	xyo	66.7	ge	93.3
nn	79.7	mu	63.8	mi	90.9	ze	84.4	do	65.9	ϕ	97.1	nn	83.7	ya	66.7	be	94.1
mu	79.9	mi	64.7	pi	90.9	be	84.8	ga	66.7	bi	100.0	mi	83.8	do	67.4	ϕ	97.0
mi	81.4	xyo	66.7	ϕ	96.5	zi	85.4	ya	69.1	na	100.0	be	86.0	xyu	68.1	na	100.0
ϕ	96.1	ϕ	90.5	na	100.0	ϕ	96.5	ϕ	90.6	bu	100.0	ϕ	96.5	ϕ	90.3	pi	100.0
	66.0		22.7		63.1		77.0		40.3		75.4		77.1		41.7		75.6

TABLE III. MODEL COMPARISON EVALUATIONS IN SECOND COMPARATIVE ANALYSIS. THE PARENTHESES INDICATE THE STANDARD DEVIATION.

machine learning model architecture	k=10, F-measure [%]	
	micro	macro
branch - CNN - unit - Transformer Encoder	93.6 (0.2)	76.5 (0.8)
(Remove P0 data)	93.9 (0.2)	77.6 (0.7)
(Add BatchNorm-1D and ReLU)	93.8 (0.2)	77.4 (1.0)
(Remove P0 data & Add BatchNorm-1D and ReLU)	94.1 (0.3)	78.4 (1.0)
branch - 2CNN - unit - Transformer Encoder	93.6 (0.3)	76.6 (1.0)
(Remove P0 data)	93.9 (0.2)	77.3 (1.0)
(Add BatchNorm-1D and ReLU)	93.5 (0.3)	76.1 (1.4)
(Remove P0 data & Add BatchNorm-1D and ReLU)	93.9 (0.2)	77.7 (0.6)

the input data for participant P0. Furthermore, in the case of adding BatchNorm-1D and ReLU, CNN-Transformer Encoder showed improvement, but no improvement was observed for 2CNN-Transformer Encoder.

Finally, we conducted a word-level accuracy evaluation of the best-performing model configuration: the branch-CNN-unit-Transformer Encoder incorporating BatchNorm-1D and ReLU. The evaluation utilized the Letter Error Rate (LER) metric, formulated in (1).

$$\mathrm{LER} = \frac{\text{Substitutions} + \text{Deletions} + \text{Insertions}}{\text{The number of characters in the reference}} \quad (1)$$

Similar to CTC, when evaluating sequences processed by removing "ϕ" tokens from the Encoder's output and merging consecutive identical fingerspelling characters, the system demonstrated strong performance with an average LER of 0.122 (0.008), indicating a low misrecognition rate.

V. DISCUSSION

A. Replacing LSTM with a Transformer Encoder

Our results demonstrated improvement in generalization performance when replacing LSTM with Transformer architectures. However, since the removal of P0 data led to improved macro F-measure scores, we cannot make strong claims about the model's ability to handle individual differences. The distinctive recognition results for P0 raise two potential explanations:

1) Whether this represents an "extreme individual difference"
2) Whether the data contains "noise" that transcends typical individual variations

During data collection, we only gathered limited participant attributes (age, gender, hearing ability, and sign language experience), which prevented us from fully analyzing the characteristics of the removed participant's data. Consequently, we cannot definitively determine whether the observed variations represent individual differences or more significant data anomalies. Thus, we have not necessarily sufficiently evaluated the machine learning model's robustness when the dataset contains outliers. Future data collection efforts should include more comprehensive participant attribute information, such as experience with JSL or Signed Japanese, to enable more thorough analysis. On the other hand, we can conclude that both CNN combination and BatchNorm-1D+ReLU application contributed to performance improvements. However, we have not yet explored the full parameter space for CNN combinations or the relationship with preprocessing parameters used in moving average calculations during dataset construction. Further comparative analysis is needed. Specifically, we plan to investigate varying the moving average calculation from 960 samples (120 sps $\times$ 8 s) to N$\times$8 samples (N sps $\times$ 8 s), along with corresponding adjustments to CNN parameters - modifying kernel size from (1, 2) to (1, N) and stride from (1, 1) to (1, N).

B. Practical use of the Transformer Encoder

Our study focused solely on the Transformer Encoder component and demonstrated its practical applicability for interface implementation. Specifically, not only did we achieve a LER of 0.122 (0.008), but we also recorded a total inference time of 0.732 s for processing the entire evaluation dataset. Given that the evaluation dataset consisted of 1,042 words, we estimated an average inference time of 0.702 ms per word.

C. Limitation

The current recognition system is not designed for real-time processing; rather, it begins recognition only after receiving a complete word input. Moreover, our continuous fingerspelling dataset consists solely of word-level data, and similar to SpellRing [26], does not include sentence-level data. Thus, we cannot evaluate continuous fingerspelling recognition at sentence-level including also NM expressions and grammatical omission. In addition, as our implementation only utilizes the Transformer Encoder, we have not conducted comparative analyses involving Transformer Decoder or CTC [28]. Moreover, our analysis does not include comparisons with other advanced architectures used in previous studies, such as Conformer [19], Spatial Temporal Graph Convolutional Networks [18], and Signformer [21].

VI. CONCLUSION AND FUTURE WORKS

In this study, we conducted a quantitative evaluation using the CNN-LSTM combined model as a baseline to assess whether the Transformer Encoder could improve Japanese fingerspelling recognition rates. Our results demonstrated that for 76 Japanese fingerspelling characters, the system achieved average micro and macro F-measures of 93.8% (0.2) and 77.4% (1.0), respectively, with a word-level LER of 0.122 (0.008). We confirmed that replacing LSTM with Transformer improved generalization performance. Future work should investigate machine learning models incorporating both Transformer Encoder and Decoder architectures. Additionally, comparative analyses including CTC and HMM approaches are necessary. Further our research will also extend to comparisons with other advanced architectures, such as Conformer [19], Spatial Temporal Graph Convolutional Networks [18], and Signformer [21]. Finally, we plan to examine the spatial characteristics differentiating fingerspelling from sign language through comparative analysis using our collected one-handed sign language dataset [36].

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Numbers JP19K11411, JP24K14243.

REFERENCES

[1] T. Tsuchiya, A. Shitara, F. Yoneyama, N. Kato, and Y. Shiraishi, "Sensor glove approach for japanese fingerspelling recognition system using convolutional neural networks," in Proceedings of The Thirteenth International Conference on Advances in Computer-Human Interactions (ACHI 2020), 2020, pp. 152–157.

[2] Y. Shiraishi, A. Shitara, F. Yoneyama, and N. Kato, "Sensor glove approach for continuous recognition of japanese fingerspelling in daily life," International Journal on Advances in Life Sciences, vol. 14, 2022, pp. 53–70.

[3] V. Ashish et al., "Attention is all you need," in Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[4] N. Mukai, N. Harada, and Y. Chang, "Japanese fingerspelling recognition based on classification tree and machine learning," in 2017 Nicograph International (NicoInt). New York, NY, USA: IEEE (Institute of Electrical and Electronics Engineers), June 2017, pp. 19–24.

[5] H. Hosoe, S. Sako, and B. Kwolek, "Recognition of jsl finger spelling using convolutional neural networks," in 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA), May 2017, pp. 85–88.

[6] M. A. Jalal, R. Chen, R. K. Moore, and L. Mihaylova, "American sign language posture understanding with deep neural networks," in 2018 21st International Conference on Information Fusion (FUSION). New York, NY, USA: IEEE (Institute of Electrical and Electronics Engineers), July 2018, pp. 573–579.

[7] M. Kondo, N. Kato, K. Fukui, and A. Okazaki, "Development and evaluation of an interactive training system for both static and dynamic fingerspelling using depth image," IEICE technical report, vol. 114, no. 512, 2015, pp. 23–28, (in Japanese).

[8] P. A. D. Bianco, O. A. Stanchi, F. M. Quiroga, F. Ronchetti, and E. Ferrante, "Signattention: On the interpretability of transformer models for sign language translation," 2024, Available: https://arxiv.org/abs/2410.14506 [retrieved: April, 2025].

[9] L. Tarrés, G. I. Gállego, A. Duarte, J. Torres, and X. Giró-i Nieto, "Sign language translation from instructional videos," in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2023, pp. 5625–5635.

[10] D. Amanda et al., "How2sign: A large-scale multimodal dataset for continuous american sign language," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2734–2743.

[11] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7784–7793.

[12] RWTH-PHOENIX-Weather 2014, "RWTH-PHOENIX-Weather 2014," 2019, URL: https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX/ [retrieved: April, 2025].

[13] RWTH-PHOENIX-Weather 2014-T, "RWTH-PHOENIX-Weather 2014-T," 2019, URL: https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX-2014-T/ [retrieved: April, 2025].

[14] C. C. Patel and P. Patel, "A comparative analysis of sign language recognition approaches across varied sign languages," in Universal Threats in Expert Applications and Solutions, V. S. Rathore, V. Piuri, R. Babo, and K. S, Eds. Singapore: Springer Nature Singapore, 2024, pp. 355–372.

[15] M. De Coster, D. Shterionov, M. Van Herreweghe, and J. Dambre, "Machine translation from signed to spoken languages: state of the art and challenges," in Universal Access in the Information Society, vol. 23. Singapore: Springer Nature Singapore, 2024, pp. 1305–1331.

[16] X. Shen, Z. Zheng, and Y. Yang, "Stepnet: Spatial-temporal part-aware network for isolated sign language recognition," ACM Trans. Multimedia Comput. Commun. Appl., vol. 20, no. 7, May 2024.

[17] N. Takayama, G. Bemitez-Garcia, and H. Takahashi, "Sign language recognition based on spatial-temporal graph convolution-transformer," Journal of the Japan Society for Precision Engineering, vol. 87, no. 12, 2021, pp. 1028–1035.

[18] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, Apr. 2018.

[19] G. Anmol et al., "Conformer: Convolution-augmented transformer for speech recognition," in Interspeech 2020, 2020, pp. 5036–5040.

[20] T. Kimura, T. Miura, and K. Kanda, "AI RECOGNITION OF JAPANESE SIGN LANGUAGE AND ITS APPLICATION," Journal of International Scientific Publications: Language, Individual & Society, vol. 18, 2024, pp. 1–10.

[21] E. Yang, "Signformer is all you need: Towards Edge AI for Sign Language," 2024, Available: https://arxiv.org/abs/2411.12901 [retrieved: April, 2025].

[22] M. E. Cabrera, J. M. Bogado, L. Fermin, R. Acuna, and D. Ralev, "Glove-based gesture recognition system," in Adaptive Mobile Robotics. World Scientific, 2012, pp. 747–753.

[23] 5DT, "5DT Data Glove 5 Ultra," 2019, URL: https://5dt.com/5dt-data-glove-ultra/ [retrieved: April, 2025].

[24] C. K. Mummadi, F. P. P. Leo, K. D. Verma, S. Kasireddy, P. M. Scholl, and K. Van Laerhoven, "Real-time embedded recognition of sign language alphabet fingerspelling in an imu-based glove," in Proceedings of the 4th International Workshop on Sensor-Based Activity Recognition and Interaction, ser. iWOAR '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 1–6.

[25] N. M. Kakoty and M. D. Sharma, "Recognition of sign language alphabets and numbers based on hand kinematics using a data glove," Procedia Computer Science, vol. 133, 2018, pp. 55–62.

[26] H. Lim et al., "Spellring: Recognizing continuous fingerspelling in american sign language using a ring," 2025, Available: https://arxiv.org/abs/2502.10830 [retrieved: April, 2025].

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, Available: https://arxiv.org/abs/1512.03385 [retrieved: April, 2025].

[28] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in Proceedings of the 23rd International Conference on Machine Learning, ser. ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 369–376.

[29] T.-W. Chong and B.-J. Kim, "American sign language recognition system using wearable sensors with deep learning approach," The Journal of the Korea Institute of Electronic Communication Sciences, vol. 15, no. 2, 2020, pp. 291–298.

[30] J. Yincheng et al., "Smartasl: "point-of-care" comprehensive asl interpreter using wearables," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 7, no. 2, Jun. 2023.

[31] R. Colin et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," J. Mach. Learn. Res., vol. 21, no. 1, Jan. 2020.

[32] L. Jiyang et al., "Signring: Continuous american sign language recognition using imu rings and virtual imu data," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 7, no. 3, Sep. 2023.

[33] C. Neidle, A. Opoku, and D. Metaxas, "Asl video corpora & sign bank: Resources available through the american sign language linguistic research project (asllrp)," 2022, Available: https://arxiv.org/abs/2201.07899 [retrieved: April, 2025].

[34] R. Takada, J. Kadomoto, and B. Shizuki, "A sensing technique for data glove using conductive fiber," in Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, ser. CHI EA '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–4.

[35] NVIDIA, "Nvidia docs hub - nvidia optimized frameworks," 2025, URL: https://docs.nvidia.com/deeplearning/frameworks/pytorch-release-notes/rel-23-11.html [retrieved: April, 2025].

[36] A. Shitara, T. Kasama, F. Yoneyama, and Y. Shiraishi, "One-handed signs: Standardization for vehicle interfaces and groundwork for automated sign language recognition," in Proceedings of The Seventeenth International Conference on Advances in Computer-Human Interactions (ACHI 2024), 2024, pp. 174–181.

CoMeSy: Multimodal Interaction with a Situated Cobot for Collaborative Tasks

Sven Milde, Alexander Jost, Rainer Blum, Jan-Torsten Milde,
Marius Schultheis, Johannes Weyel, Tobias Müller, Thies Beinke
Niklas Schreiner, Julian Heumüller, Dennis Möller, Frank Hartmann
Department of Computer Science
Fulda University of Applied Sciences
Fulda, Germany
e-mail: (sven.milde, alexander.jost, rainer.blum)@cs.hs-fulda.de, milde@hs-fulda.de,
(marius.schultheis, johannes.weyel, tobias.mueller, thies.beinke, niklas.schreiner)@et.hs-fulda.de,
(julian.heumueller, dennis.möller, frank.hartmann)@alpaka-innovation

***Abstract*—The CoMeSy project is developing a system for multimodal interaction between humans and cobots, where the cobot acts as an intelligent assistant. The system uses speech and gestures as input, and responds with speech, sounds, actions, and visual feedback. A key challenge is dynamically creating action plans based on human input, world knowledge, and visual perception. The system integrates several technologies, including speech recognition and synthesis, image processing, object detection, hand tracking, and acoustic feedback. Currently in development, the project aims to address intelligent communication, situational understanding, dynamic planning, reactive behavior, and robust handling of interruptions, with plans for empirical evaluation.**

Keywords-Multimodal interaction; Human-Robot Interaction (HRI); Collaborative robotics; Cobot as an intelligent assistant.

I. Introduction

Collaborative robotics, especially the interaction between humans and robots (HRI), has become a central research area in robotics [1], [2]. The CoMeSy (CoMeSy: *Co*bot *Me*nsch *S*ymbiose, German for Cobot Human Symbiosis) project focuses on multimodal, situationally embedded interaction with a cobot with the aim of jointly carrying out action processes. The cobot should take on the role of an intelligent, supportive assistant; the human takes the lead in the process. The interaction between human and cobot is realized through multimodal inputs (speech and gestures) [3]. The cobot also reacts multimodally via speech and acoustic signals, can perform situation-adapted actions and give visual feedback via a table projection. A central challenge consists in the creation of dynamic action plans, which are based on the instructions of the human, the inherent world knowledge of the current domain as well as the (visual) perception of the current state of the world.

The rest of the paper ist structerd as follows: Section II will first provide an overview of related work, followed by an explanation of the underlying research agenda in Section III. Section IV will elaborate on the basic technical scenario, for which the situated action control will be described in Section V. Section VI concludes the paper and summarizes the central findings.

II. Related Work

The development of multimodal human-robot interaction (HRI) has seen significant progress in recent years, particularly with regard to the integration of speech, gestures and other modalities. The objective of this integration is to establish seamless and intuitive communication between humans and machines.

A comprehensive overview of human-robot collaboration is provided in [4]–[6]. The impact of robot movements on human-robot collaboration is analysed in [7] and is pertinent to the planning of robot movements to avoid collisions with the user.

The significance of multimodal interfaces is highlighted in [8], with a primary focus on communication through speech and gesture. In particular, it is emphasised that the content of the communication is more important than the method, which is facilitated by the use of multimodality. This assertion is further substantiated by comparative analyses of unimodal and multimodal user interfaces in [9], [10]. The findings of these studies demonstrate that multimodality serves to reduce cognitive load. This is considered to be of crucial importance for the system concept under discussion, with the result that the user can concentrate primarily on the work task and act intuitively with the robot. In a similar vein, the study by Turk [11] corroborates the notion that multimodality fosters enhanced user acceptance. Common misconceptions in the domain of multimodal interaction and its advantages are illustrated by Oviatt [12], while Baltruvsaitis et.al. [13] examines the relationships between the modalities and describes various forms of fusion, as well as the challenges that arise when evaluating multimodal interaction. The challenges of evaluating multimodal interaction are also considered in the context of this work, as it is essential to interpret the various modalities correctly and to integrate them to form a comprehensive understanding.

The integration of Large Language Models (LLMs) in robot planning is investigated in [14], wherein the capacity of LLMs to generate Behaviour Trees for robot task planning is analysed. An interactive planning approach with LLMs that enables agents to handle multiple tasks in open environments

is presented in Wang et.al. [15]. In this work, however, the LLM will be used to transition from a voice command to a series of granular actions. Additionally, Ao et al. [16] adopt an approach integrating speech and gestures into a planning method with an LLM, with the objective of generating a sequence of actions for a robot. Another LLM-driven approach to robust task planning and execution is presented in [17], where mainly visual information is used to obtain a worldview and plan tasks in it, whereas in this work a multimodal input is used.

The subject of computer-based hand gesture recognition for HRI is addressed in depth in the publication by Qi et al. [18]. The utilisation of direct gestures for communication with robotic systems is addressed in [19]. A context-aware robotic assistance system based on pointing-based gestures to support humans is described in [20] and demonstrates the significance of the robot's world knowledge for this work, ensuring the accurate interpretation of gestures.

A comprehensive overview of Behaviour Trees (BTs) in robotics and artificial intelligence is given in [21], [22], highlighting the application of BTs to control robot behaviour as it will be used in this work. A key benefit of behavior trees is their capacity for dynamic response to alterations during runtime.

III. Research agenda: an intelligent, collaborative system

The primary goal of the project is the development of an intelligent, collaborative system that can be controlled by natural language communication and intuitive gestures. CoMeSy should be able to perform complex tasks in dynamic environments and flexibly adapt to unexpected situations. The research agenda includes the following points:

1) *Intelligent communicative behavior within the framework of collaborative action:* the system should be able to understand and generate natural language and effectively integrate it into the cooperative work process. To do this, the system must be able to exchange relevant information, give and receive instructions, provide and process feedback, and coordinate its own actions with the human actor.
2) *Situational interpretation of linguistic instructions and gestures*: CoMeSy should interpret linguistic instructions and gestures situationally, i.e., grasp the meaning of language and gestures not in isolation, but in the context of the respective situation. This includes resolving ambiguities, processing ellipses, deictic expressions, interjections, and prosodic markings, accepting the pragmatic use of language, and ideally recognizing the intention of the communication partner.
3) *Dynamic action planning based on a linguistic target specification*: the system should be able to create dynamic action plans based on a linguistic target specification. This requires the ability to break down complex tasks into subtasks, develop suitable action strategies, and adapt them flexibly to changing world states
4) *Integration of reactive behavior control into action execution*: the system is able to react quickly and appropriately to unexpected events and changes in the environment and, if necessary, adapt action plans and develop alternative strategies.
5) *Robust system behavior in case of interruptions or dynamic changes of the world state*: CoMeSy should demonstrate robust system behavior in case of interruptions (such as communication breakdowns) or dynamic changes in the world state.

IV. Scenario: collaborative work

CoMeSy is currently in its first construction phase. In addition to the physical construction of a work cell, the components for sensor data preprocessing, in particular, have been implemented:

- Speech recognition and speech synthesis (whisper.cpp, coqui/XTTS-2)
- Reading out the current camera images (opencv)
- Visual object detection and classification (yolo11m with fine-tuning via LabelStudio)
- Hand detection and posture analysis (mediapipe and tensorflow-based ANN (Artificial Neural Network)
- Synthesis of acoustic feedback signals (SuperCollider)

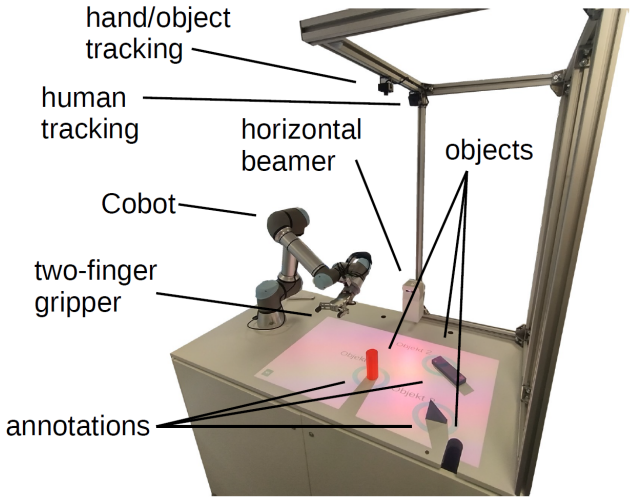

Figure 1. The work cell: Cobot and human share a workspace. Two cameras capture the workspace and the human. Visual information can be displayed on the work surface via a horizontal beamer.

A. The work cell

Human and Cobot (here Universal Robot UR-5, or IGUS Rebel) work together in a work cell (see Fig. 1) and share a common work surface. The UR-5 is equipped with a 2-finger gripper that allows it to grasp and place objects. The 2-finger gripper is also equipped with a force-torque sensor that enables precise measurements of the forces and torques acting on the gripper.

A depth image camera, which is mounted above the Tool Center Point (TCP) of the robot, captures both the objects to be gripped and the current position of the 2-finger gripper. The captured objects are initially designed in simple geometric shapes, such as cuboids, cylinders, prisms and cubes, whereby they can be in different colors.

The work surface itself is captured by a vertically aligned camera. This camera is able to identify the position and orientation of the objects on the work surface. In addition, it has the ability to identify the hands of the human, to distinguish between the left and right hand and to determine the respective hand posture (see Fig. 2).

The upper body of the human is captured by an inclined camera to obtain additional information about the posture and movements of the user. A short-range beamer projects an image onto the work surface to provide visual information or instructions.

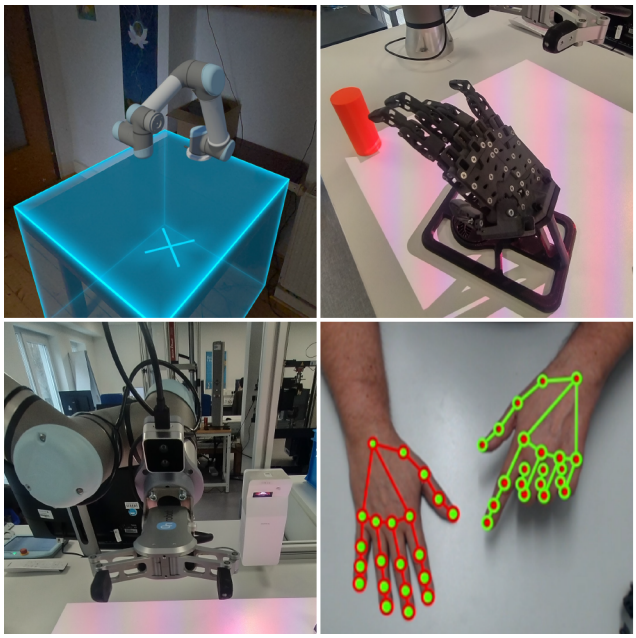

Figure 2. Impressions: a): The prototype of the AR application for the UR-5. b): Prototype 5-finger hand c): View of the 2-finger gripper, depth camera and short-range beamer. d): Hand posture recognition

The multimodal interaction is completed by acoustic inputs via a 4-channel directional microphone and acoustic outputs via a loudspeaker. This setup enables a comprehensive recording of the interactions between human and cobot as well as an intuitive and versatile communication interface.

B. System architecture

The system architecture (see Fig. 3) consists of two major subcomponents. The first consists of ROS2 (Robot Operating System version 2) nodes, which capture the various sensor data and, in particular, perform visual pre-processing and recognition. There are also nodes for controlling the overall system, which relate to different modalities, such as speech, audio, or movement. The Blackboard serves as a memory in this subcomponent and records all sensor data and intermediate results. The BehaviorTree [22] is the central control of the system, which ensures a reactive and dynamic behavior of the robot.

The second sub-component consists of various parallel processes, some of which are particularly computationally intensive and therefore have been outsourced to a corresponding AI server. This includes, above all, audio processing with speech recognition, keyword spotting, and audio synthesis. The integration of a Large-Language Model (LLM) to evaluate speech commands is also implemented as a seperate process. For this purpose, llama3.3 is run locally in Ollama [23]. Information about the internal state of the system, such as recognized objects or understood commands, is visualized on the work surface via a short-range beamer. This visualization is also implemented as a seperate process. The communication between the two sub-components is implemented as a REST API via FastAPI [24].

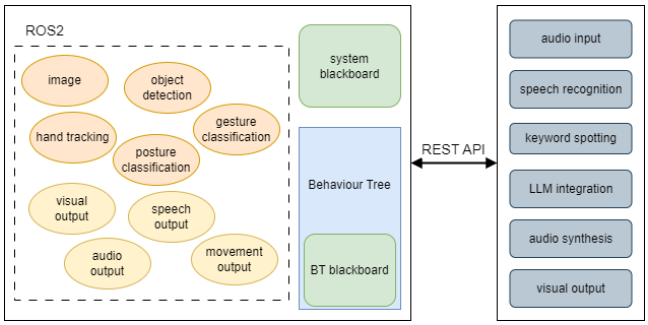

Figure 3. The system architecture: ROS2 nodes control the robot actions. Computationally expensive processes run on the AI (Artificial Intelligence) Server. The two parts communicate via a REST API.

V. Situational Action Control

A. BehaviorTrees

In order for the system to be able to dynamically react to user input and sensor readings at any time, the individual running processes must be fine-grained and interruptible. With the help of BehaviorTrees, this property can be achieved, since BehaviorTrees are run at a defined frequency and the individual nodes are stopped when they receive a corresponding signal from higher levels of the tree. This signal can be generated by other nodes or reactively when data on the Blackboard has been changed."

To solve varying tasks using the BehaviorTree [22], the individual actions that the robot is able to perform are represented in the smallest possible subtrees. Based on user input, a sequence of granular actions can then be defined, which can then be dynamically loaded into the tree as a sequence. This allows different sequences of operations to be put together which are not pre-programmed into the system as predetermined fixed processes. The basic BehaviorTree (see Fig. 4) thus consists of a node that calibrates the entire system when it is started and performs a system check, a higher-level

error handling, and a node that dynamically loads the various subtrees, depending on the command given.

In addition, a parallel node should run as high as possible in the BehaviorTree, ensuring the processing of the sensor data. For this purpose, different topics are subscribed within the ROS network and stored partly directly and partly after prior processing in the Blackboard, so that the BehaviorTree can access them quickly and easily.

The use of the Blackboard has the advantage that during a single tick within the BehaviorTree, the data in the Blackboard is not changed, so that the entire BehaviorTree works with the same consistent values. If individual nodes were to subscribe to the same topics independently, the values could change and the tree could thus work with inconsistent data.

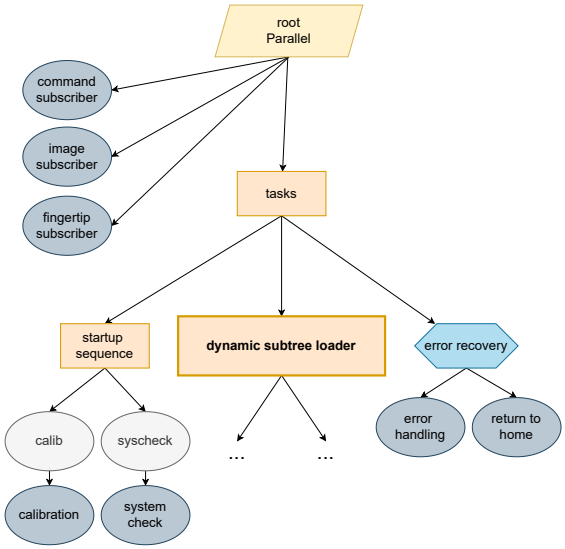

Figure 4. The basic BehaviorTree: At the very top (marked yellow) is the parallel node, which contains the various subscribers and the main task.

B. Multimodal action sequences: examples

The interaction between humans and robots in collaborative robotics requires intuitive and efficient communication strategies. Speech-gesture-control allows humans to interact with the robot in a natural intuitive way and it also allows to coordinate complex tasks. In the context of a work cell, where objects are located on a table, the situation-based formulation of action instructions is crucial for the success of human-robot collaboration.

In the following, 3 exemplary multimodal interaction sequences are outlined to illustrate the complexity of speech-gesture controlled human-robot interaction. We assume the following initial situation: several geometric objects are positioned on the table. The cobot is ready to take on tasks.

A crucial element for the interaction between humans and robots is *visual perception*. The camera enables the cobot to perceive its environment, recognize objects, and move safely within the space. In the example dialogue in table I, objects are verbally referenced via shape, color, and spatial relation, and identified through the visual channel.

TABLE I. ACTION-SEQUENCE 1: COLLABORATIVE BUILDING

agent	action/utterance
Human	"Hand me the red cube."
Cobot	(Moves camera to the tabletop, identifies the red cube.) "Alright, which arm would you like?"
Human	(Holds out left hand.) "Left, please."
Cobot	(Grasps the cube with the left hand, lifts it, and rotates it so that the human can easily grasp it.)
Human	(Grasps the cube and places it on top of the blue pyramid.)
Cobot	(Observes the action.) "Would you like to continue building?"

TABLE II. ACTION-SEQUENCE 2: FAILED COMMUNICATION

agent	action/utterance
Human	(Points at the red cube and makes a rotating motion with their hand.) "Turn the cube."
Cobot	(Misinterprets the gesture.) "Would you like me to rotate the cube around its own axis?"
Human	(Shakes head.) "No, turn it so that the red side faces upwards."
Cobot	(Re-analyzes the situation.) "Ah, now I understand." (Rotates the cube accordingly.)

In the situational embedding of the cobot, it is assumed that communication and action can *fail*, and in many cases they will. Errors can occur due to misunderstandings (see Table II) or unforeseen events (see Table III). It is, therefore, important that humans have the opportunity to intervene in the work process and make dynamic corrections if necessary.

VI. SITUATION-BASED LANGUAGE CONTROL

Language processing in CoMeSy takes place on two conceptual levels. At the lower level 1, we consider strictly situationally embedded instructions, which often have deictic constructions or are situationally expanded and interpreted as linguistic ellipses. These instructions usually refer to the current running sub-action and allow humans to linguistically control and modify it.

At the higher level 2, the specification of more complex actions takes place; actions that are constructed as a sequence of sub-actions and that need to be planned in a dynamic and situatiated way.

The following are examples of level 1 instructions, explained in detail. We differentiate between *general instructions*, *position-related instructions*, *manipulation-related instructions*, and *situational modifiers*.

A. general instructions

- *Stop, Halt*: Interrupt the current sub-action.
- *Continue*: Resume the previously interrupted sub-action.
- *Yes, No*: Confirm or reject instructions or feedback from the robot.
- *Faster, Slower*: Adjust the robot's action speed (to suit human needs).
- *Correct, Incorrect*: Evaluate the robot's action, allowing for error correction.

B. position-related instructions

- *Left, Right, Forward, Backward, Up, Down*: Description of directions and positions in space, often in combination with prepositional phrases (e.g., "in front of the object").
- *In front of, Behind*: Describe the position of an object in spatial relation to one another.
- *Above, Below*: Also relational, but refer to the vertical positioning of objects.

C. manipulation-related instructions

- *Grasp, Release*: Describe the basic actions of the robot gripper.
- *Turn/Rotate the hand, The other way around, Clockwise, Counterclockwise*: Control of the robot's gripper during rotational movements.

TABLE III. ACTION-SEQUENCE 3: UNFORSEEN EVENTS

agent	action/utterance
Human	"Can you bring the green sphere to the pyramid?"
Cobot	"Sure!" (Moves to the sphere, grasps it.)
-	(Unexpected event): The sphere rolls away and falls off the table.
Cobot	(Searches for the sphere with the camera.) "Oh, the sphere fell. I'll try to get it." (Moves to the floor, grasps the sphere.)
Human	(Points to a new spot on the table.) "Put it there."
Cobot	(Places the sphere in the desired location.)

D. situational modifiers

- *A little more, A little bit further* : Fine-tuning of positions and movements. These instructions have lower precision and may require additional visual or tactile feedback.

VII. CONCLUSIONS

CoMeSy is currently in the technological development phase. Hardware and software components have been identified, a system architecture has been outlined, and in the following weeks, with the completion of the first iteration, integration into a working prototype will take place. Subsequently, the points of the research agenda will be addressed. In addition to carrying out the necessary functional tests, the system will be empirically evaluated via subject tests. Furthermore, the complexity of the scenario is to be gradually increased, for example, by integrating the 5-finger hand and using further objects.

REFERENCES

[1] R. Müller *et al.*, *Handbuch Mensch-Roboter-Kollaboration*. Carl Hanser Verlag GmbH Co KG, 2023.

[2] H. Su *et al.*, "Recent advancements in multimodal human–robot interaction," *Frontiers in Neurorobotics*, vol. 17, p. 1 084 000, 2023.

[3] R. Younes, F. Elisei, D. Pellier, and G. Bailly, "Impact of verbal instructions and deictic gestures of a cobot on the performance of human coworkers," in *2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*, IEEE, 2024, pp. 1040–1047.

[4] B. Chandrasekaran and J. M. Conrad, "Human-robot collaboration: A survey," in *SoutheastCon 2015*, IEEE, 2015, pp. 1–8.

[5] M. Peshkin and J. E. Colgate, "Cobots," *Industrial Robot: An International Journal*, vol. 26, no. 5, pp. 335–341, 1999.

[6] A. Bauer, D. Wollherr, and M. Buss, "Human–robot collaboration: A survey," *International Journal of Humanoid Robotics*, vol. 5, no. 01, pp. 47–66, 2008.

[7] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa, "Effects of robot motion on human-robot collaboration," in *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, 2015, pp. 51–58.

[8] D. Perzanowski, A. C. Schultz, W. Adams, E. Marsh, and M. Bugajska, "Building a multimodal human-robot interface," *IEEE intelligent systems*, vol. 16, no. 1, pp. 16–21, 2001.

[9] S. Oviatt, "Advances in robust multimodal interface design," *IEEE computer graphics and applications*, vol. 23, no. 05, pp. 62–68, 2003.

[10] S. Oviatt, R. Coulston, and R. Lunsford, "When do we interact multimodally? cognitive load and multimodal communication patterns," in *Proceedings of the 6th international conference on Multimodal interfaces*, 2004, pp. 129–136.

[11] M. Turk, "Multimodal interaction: A review," *Pattern recognition letters*, vol. 36, pp. 189–195, 2014.

[12] S. Oviatt, "Ten myths of multimodal interaction," *Communications of the ACM*, vol. 42, no. 11, pp. 74–81, 1999.

[13] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[14] Y. Lai *et al.*, "Nmm-hri: Natural multi-modal human-robot interaction with voice and deictic posture via large language model," *arXiv preprint arXiv:2501.00785*, 2025.

[15] Z. Wang *et al.*, "Describe, explain, plan and select: Interactive planning with llms enables open-world multi-task agents," *Advances in Neural Information Processing Systems*, vol. 36, pp. 34 153–34 189, 2023.

[16] J. Ao, F. Wu, Y. Wu, A. Swikir, and S. Haddadin, "Llm as bt-planner: Leveraging llms for behavior tree generation in robot task planning," *arXiv preprint arXiv:2409.10444*, 2024.

[17] C. Rivera *et al.*, "Conceptagent: Llm-driven precondition grounding and tree search for robust task planning and execution," *arXiv preprint arXiv:2410.06108*, 2024.

[18] J. Qi, L. Ma, Z. Cui, and Y. Yu, "Computer vision-based hand gesture recognition for human-robot interaction: A review," *Complex & Intelligent Systems*, vol. 10, no. 1, pp. 1581–1606, 2024.

[19] M. Pascher *et al.*, "Hands-on robotics: Enabling communication through direct gesture control," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 822–827.

[20] S. Drolshagen, M. Pfingsthorn, and A. Hein, "Context-aware robotic assistive system: Robotic pointing gesture-based assistance for people with disabilities in sheltered workshops," *Robotics*, vol. 12, no. 5, p. 132, 2023.

[21] M. Iovino, E. Scukins, J. Styrud, P. Ögren, and C. Smith, "A survey of behavior trees in robotics and ai," *Robotics and Autonomous Systems*, vol. 154, p. 104 096, 2022.

[22] M. Colledanchise and P. Ögren, *Behavior trees in robotics and AI: An introduction*. CRC Press, 2018.

[23] Software: ollama, last access: May, 2025 https://github.com/ollama/ollama.

[24] Software: FastAPI last access: May, 2025 https://github.com/fastapi/fastapi.

Vision-Controlled Hand Gesture Recognition System for Home Automation

Luis Gilberto Rangel Pérez
Centro Universitario de los Valles
Universidad de Guadalajara
Ameca, Jalisco, México
Email: luis.rangel4328@alumnos.udg.mx

Rodolfo Omar Domínguez García
Centro Universitario de los Valles
Universidad de Guadalajara
Ameca, Jalisco, México
Email: odomi@academicos.udg.mx

Miriam González Dueñas
Centro Universitario de los Valles
Universidad de Guadalajara
Ameca, Jalisco, México
Email: miriam.gduenas@academicos.udg.mx

Miguel Angel de la Torre Gómora
Centro Universitario de los Valles
Universidad de Guadalajara
Ameca, Jalisco, México
Email: miguel.dgomora@academicos.udg.mx

***Abstract*— The increasing need to support elderly individuals and people with disabilities has driven the development of innovative assistive technologies that enhance independence and improve quality of life. This manuscript presents a gesture-controlled home automation system that allows users to intuitively operate household devices, such as lights and blinds, without physical contact. The architecture of the proposed system is supported by a Raspberry_Pi 4, with OpenCV and MediaPipe for real-time hand gesture recognition. The system interprets commands and executes corresponding actions on connected devices, taking advantage of finger positions. Unlike traditional automation solutions that rely on voice commands or touch interfaces, this approach offers an accessible and hygienic alternative, particularly beneficial for individuals with mobility or speech impairments. The system exhibits high accuracy, rapid response times, and user-friendly operation, making it a cost-effective and scalable solution for smart home applications. This is achieved through the integration of low-cost devices and open-source libraries. Through this work, we contribute to the advancement of inclusive and contactless technology, fostering greater autonomy and accessibility in everyday living environments.**

Keywords- Assistive technology, hand gesture recognition, home automation, Raspberry_Pi, accessibility.

I. INTRODUCTION

Jesse Jackson, a civil rights activist, said that "inclusion is the key to growth". For this reason, Assistive Technology (AT) is being increasingly promoted worldwide, as it encompasses equipment, devices, tools, and software designed to reduce barriers, enhance accessibility, and improve access to resources. These technologies empower individuals with hearing impairments, visual impairments, physical disabilities, or progressive loss of autonomy, fostering greater independence, self-determination, and inclusion in daily life [1]. According to the World Health Organization (WHO), there are 1 billion people in the world who need an assistive device, yet only 1 in 10 has access to one. In low- and middle-income countries, only 5% to 15% have access to AT. In Mexico, the 2020 National Census of Population and Housing shows that 20.8 million people - 16.5% of the total population - have some form of disability [2].

To enhance quality of life, the WHO, together with the Global Cooperation in Assistive Technology (GATE) initiative, recognizes access to assistive technologies as a universal right. These technologies play a crucial role in enabling individuals to lead healthy, productive, independent, and dignified lives, facilitating their participation in education, the workforce, and social activities.

AT, serving individuals with disabilities, has significantly contributed to their health and well-being, as well as that of their families. It not only helps prevent the deterioration of their condition but also plays a crucial role in reducing public health expenditures [2]. AT enables and promotes inclusion and participation, especially for people with disabilities, older people and people with non-communicable diseases. The main purpose of these products is to maintain or enhance people's function and autonomy, thus promoting their well-being. They enable people to live a dignified, healthy, productive and independent life, and to learn, work and participate in social life [3].

Currently, one billion people require an assistive device, and this number is projected to exceed two billion by 2030, highlighting the growing global demand of AT. Although anyone can need an assistive product at some point in their lives, the most common users are adults and children with disabilities, older people and people with chronic conditions, such as diabetes and dementia. Examples of assistive or supportive products include hearing aids, wheelchairs, eyeglasses, prostheses and memory aids, among many others. In addition to promoting autonomy and well-being, these products can help prevent or reduce the impact of secondary conditions, such as lower limb amputation in diabetics. They can also reduce the need for and impact of dependency on caregivers and medical and support services. In addition, access to appropriate assistive devices can have

 ISBN: 978-1-68558-268-5

major benefits for community development and economic growth [4].

Despite the global need for and recognized benefits of assistive or supportive products, access to them remains limited. Addressing this unmet need is essential to progress towards the Sustainable Development Goals and the implementation of the Convention on the Rights of Persons with Disabilities [5][6].

It also responds to the growing demand for home automation solutions that are accessible and adaptable to the needs of different users. Unlike traditional home automation systems that require touch interfaces or voice commands, gesture-based control offers an innovative alternative that is ideal for people who have difficulty speaking or using physical devices. The COVID-19 pandemic highlighted the importance of non-contact technologies to reduce the spread of disease, and these types of solutions can help improve hygiene and safety in different contexts.

Finally, this project has significant educational and technological value as it combines the use of open-source hardware and software, promoting learning in areas such as computer vision, image processing and embedded control. Its modular and scalable design allows for future improvements and adaptations, making it a basis for the development of more complex automation and accessibility systems.

The project was motivated by the need to help the elderly or people with reduced mobility, who often find it difficult to perform everyday tasks such as turning lights on and off or manipulating objects at different heights. With the advancement of technology, it is possible to develop innovative solutions that improve their quality of life, providing greater autonomy and comfort at home [7].

The present manuscript presents the design and implementation of a control system based on hand gestures. The system enables intuitive and contactless control of light bulbs and the opening/closing of blinds, enhancing accessibility and ease of use. A Raspberry_Pi 4 is used as a central processing unit, together with the OpenCV [8] and MediaPipe libraries [9] for real-time gesture recognition and analysis. Through finger position recognition, the system interprets various commands and performs specific actions on the connected devices, providing an efficient and accessible user experience. The Raspberry_Pi 4 was chosen for its processing capacities, low power consumption, and versatility in integrating with various sensors and actuators. In addition, its compatibility with Python [10] and specialized computer vision libraries, such as OpenCV and MediaPipe, allows the development of optimized algorithms for real-time image detection and processing. These elements ensure accurate and reliable operation, even in environments with lighting variations or unconventional hand positions.

In addition to its application in the home, this project has the potential to be extended to other areas such as industrial automation, advanced domotics, and accessibility systems for people with disabilities. The ability to control devices through gestures without the need for physical contact represents a safe and hygienic alternative, especially in environments where interaction with surfaces must be minimized, such as hospitals or laboratories.

In conclusion, this system aims to provide an innovative, accessible and efficient solution for home automation, improving the independence and quality of life of users. With advanced computer vision and embedded processing technologies, new possibilities are explored in the field of contactless control, with potential applications in various areas of daily life.

The rest of this paper is organized as follows. Section II describes the related work. Section III describes the materials and methodology used in the implementation of the proposed system. Section IV presents and discusses the results, and Section V summarizes the conclusions and future works.

II. RELATED WORK

The interest of the research community in AT is evidenced in a few recent works. For instance, in [11], the current populace of the elderly is apparently abandoned by the younger generations due to their individual circumstances. To enhance vitality and improve the well-being of elderly individuals, an assisted home care system can serve as a valuable solution by offering comprehensive nursing care and continuous monitoring. The proposed system used voice and gesture (MPU6050 accelerometer) to control home appliances like turning on/off the light, closing/opening of curtains, TV, and fan or AC within the living spaces. The system also monitors real-time activities like heart rate and body temperature for the elderly citizens. In the event of an emergency, such as anomalous behaviors indicating a stroke, the proposed system automatically triggers an alarm and turns on an emergency light to alert family members or caregivers. This smart environment can set the temperature and help control the living parameters based on the users' comfort and their health conditions [11]. Gourob et al. [12] reveal that Human-Robot Interaction (HRI) has become an important topic in today's robotic world, especially in assistive robotics. Vision based hand recognition systems provide solutions for these types of human demands. In the system proposed in [12], users do not need to physically operate any devices. Instead, a camera captures hand movements, and these recordings serve as input for gesture-based control, enabling seamless and contactless interaction with the system. Regarding the elderly people, patients and disabled people can benefit from this kind of gesture control. Besides, AT can be used in extreme environments where direct contact is impossible or impractical. Here, a vision-based hand gesture system for controlling robotic hands has been developed. The main intention in [12] is to implement a vision-based gesture recognition system that recognizes data gathered from hand movements through a camera. In [13], the authors reviewed the sign language research in the vision-based hand gesture recognition system from 2014 to 2020. They identified progress and relevant needs that require more attention. The review shows that vision-based hand gesture recognition research is an active field of research, with many studies conducted, resulting in dozens of articles published annually in journals and conference proceedings. Most of the articles focus on three critical aspects of the vision-based hand gesture recognition system, namely: data acquisition, data

environment, and hand gesture representation. They also reviewed the performance of the vision-based hand gesture recognition system in terms of recognition accuracy. For the signer dependent, the recognition accuracy ranges from 69% to 98%, with an average of 88.8% among the selected studies. The lack in the progress of continuous gesture recognition could indicate that more work is needed towards a practical vision-based gesture recognition system [13].

The authors in [14] propose a Home Automation System with hand gesture recognition based on Mediapipe, using Arduino UNO. As one gets older, his/her mobility tends to decrease. Therefore, simple tasks such as getting up to switch the lights on or turning the fan off can become difficult. Thus, it became imperative to create a system which allows them to perform these tasks - a “Hand Recognition based Home Automation System”. Various methods for gesture recognition have been discussed as well as how every model produces a varying training time and accuracy. Although an average accuracy was found to be 98% with the majority of the Machine Learning (ML) models and MediaPipe's technology, our suggested methodology demonstrates that MediaPipe with Dense may be effectively utilized as a tool to correctly recognize complicated hand gestures. Additionally, training this particular model takes much less time compared to the other models. Faster real-time detection of gestures demonstrates the efficiency of the model. In [15], the authors propose a remote-control method based on one-handed gestures for mobile manipulator, so that the operator can control the entire robotic system with only one hand. In this study, they combined real-time hand key points detection technology provided by MediaPipe, with the RealSense D435i depth camera to address the inaccuracy in depth recognition problem of the original method. Then, the position, pitch, and rotation of the hand are analyzed to generate control commands. A lightweight gesture recognition model based on a Gated Recurrent Unit (GRU) is proposed to use specific gestures for switching between controlled objects.

In [16], the authors focus on the design of a gesture-controlled robotic arm that is navigated with the help of a webcam and OpenCV-enabled real-time hand tracking. Embedded with OpenCV’s hand tracking module, the system successfully identifies the hand landmarks from the live camera captured hand movement of the user. The state of each finger is represented as a string that includes the $ symbol and a number, such as $00010. This is transmitted to an Arduino UNO board, and, based on this information, the

robotic arm can move its fingers up or down. Control of the flows of the movement of the robotic arm is done using the Arduino UNO whereby the servo motors receive signals to move 0 or 180 degrees. This is mainly due to the efficiency, precision, and short latency of the suggested system, which makes it novel. It gives the user a simple graphics interface to control robotic arms without the necessity to know programming or have specific hardware. This technology provides an effective solution in robotic manipulation and presents the enormous potential to enhance HRI in different application areas [16].

III. MATERIALS AND METODS

Gesture recognition is a technique that interprets hand or body movements to interact with electronic devices without the need for physical contact. It has a wide range of applications, from video games and augmented reality to home automation and accessibility for people with disabilities.

A. Raspberry_Pi 4 and its features

The Raspberry_Pi 4 [17] is a low-cost, high-performance microcomputer ideal for image processing and embedded control applications. Its key features include:

- Quad-Core ARM Cortex-A72 processor.
- Up to 8 GB of RAM.
- General Purpose Input Output (GPIO) ports for connecting to sensors and actuators.
- Support for Python and computer vision libraries.

Raspberry_Pi 4 is used to:

1. Capture real-time video from a Pi or USB camera.
2. Process the images using OpenCV and MediaPipe to recognized gestures.
3. Send signals via the GPIO ports to the actuators that control, for example, spotlights and blinds.

In this system, GPIO ports are used to switch spotlights on and off using relays or voltage control modules and control a motor to open and close blinds based on detected gestures. Communication between the software and the actuators is achieved by programming the Raspberry_Pi 4 in Python, allowing for easy and efficient integration. This can be seen in Figures 1 and 2.

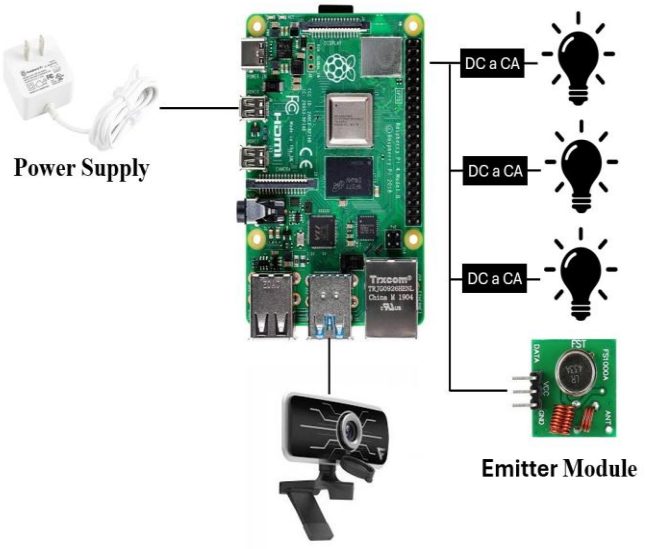

Figure 1. Schematic diagram.

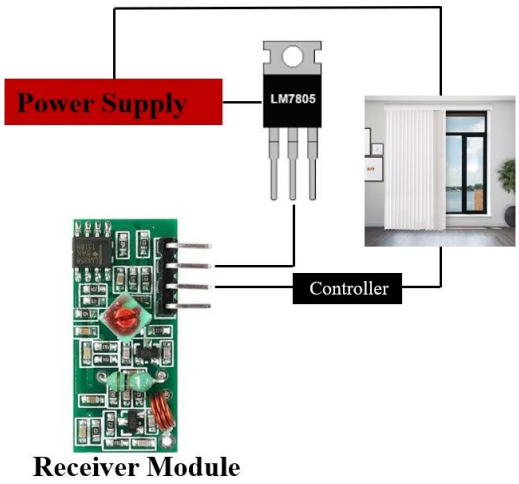

Figure 2. Module actuator blinds.

Figures 3 and 4 show the prototype implemented to carry out the tests and observe how the programmed gestures interact with the system. With the corresponding gesture, the lights are turned on or off accordingly. With other gestures, the blinds are closed or opened (the motor turns on or off).

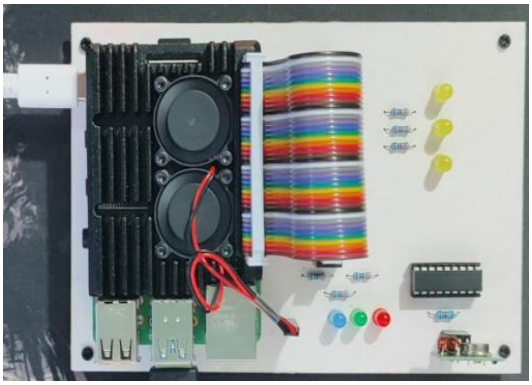

Figure 3. Prototype based on Raspberry_Pi 4B+.

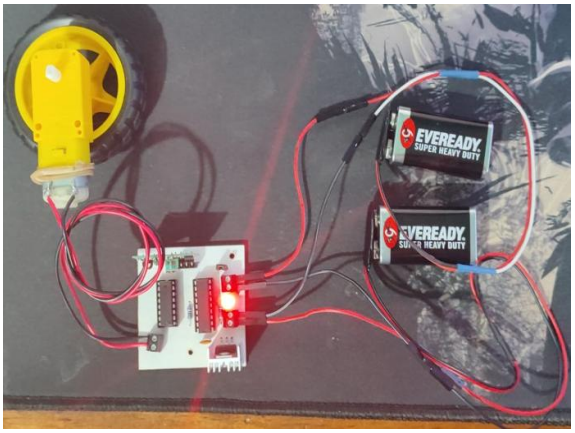

Figure 4. Prototype actuator for blinds.

B. Gesture recognition techniques

There are several techniques for recognizing and interpreting gestures, of which the following stand out:

- Sensor-based: These use devices such as accelerometers, gyroscopes and infrared sensors to detect movement.
- Computer vision-based: They use cameras and image processing algorithms to identify the position and movement of the hand.
- Hybrid: Combining physical sensors with computer vision for greater accuracy.
- This project: uses a technique based on computer vision because it enables gesture recognition without the need for additional physical devices.
- Two key libraries are used to implement gesture recognition in this project:
  OpenCV (Open-Source Computer Vision Library): It provides tools for real-time image acquisition, processing and analysis.
  MediaPipe: A framework developed by Google that uses artificial intelligence and machine learning to efficiently recognize hands, faces and poses. MediaPipe provides pre-trained hand recognition models that enable segmentation and analysis of finger movements with high accuracy.

The ability to perceive the shape and motion of hands can be a vital component in improving the user experience across a variety of technological domains and platforms. While coming naturally to people, robust real-time hand perception is a decidedly challenging computer vision task, as hands often occlude themselves or each other (e.g., finger/palm occlusions and handshakes) and lack high contrast patterns [9].

MediaPipe Hands is a high-fidelity hand and finger tracking solution. It employs Machine Learning (ML) to infer 21 3D landmarks of a hand from just a single frame. Whereas current state-of-the-art approaches rely primarily on powerful desktop environments for inference, our method achieves real-time performance on a mobile phone, and even scales to multiple hands [9].

After the palm detection over the whole image, our subsequent hand landmark model performs precise keypoint localization of 21 3D hand-knuckle coordinates inside the detected hand regions via regression, that is direct coordinate prediction, as shown in Figure 5.

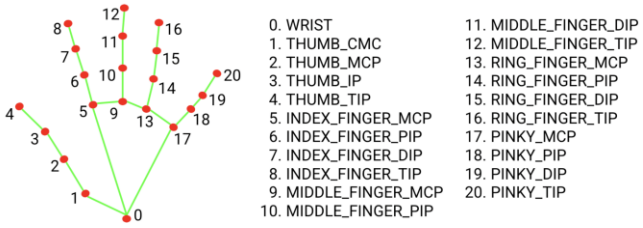

Figure 5. Hand landmarks (taken from [9]).

Figure 6 shows examples of hand gestures that MediaPipe and its training model are able to detect and generate [9].

Figure 6. Examples of hand gestures (taken from [9]).

According to Table I, gesture A is the first action to set the *system enable* state to recognize the on and off orders of the devices. This leads to gestures B or C, which indicate that a device wants to be turned on or off. The system then waits for any gesture from D to H to turn on or off the chosen device. Once the system receives gesture from D to H, and the required action is given, the system returns to the *system enable* state, waiting for gesture A in order to repeat the process.

TABLE I. MENU OF OPTIONS

Gesture	Thumb	Index	Middle	Ring	Pinky	Action
A	0	0	0	0	0	System enable state
B	1	1	1	1	1	It enters you into the power on menu
C	1	0	0	0	1	It takes you to the shutdown menu
D	1	0	0	0	0	Spotlight on/off 1
E	1	1	0	0	0	Spotlight on/off 2
F	1	1	0	0	1	Spotlight on/off 3
G	0	1	0	0	1	Motor (blinds)

IV. RESULTS

The prototype was tested in a controlled environment to observe its operation and estimate its accuracy, response time and ease of use. Figure 7 shows the system enable state (gesture A).

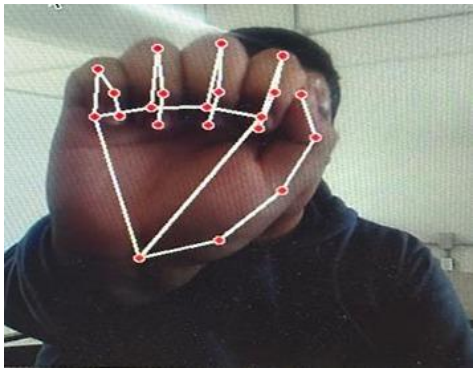

Figure 7. Gesture A, *system enable state*.

Figure 8 shows gesture B, which initiates the power-on menu.

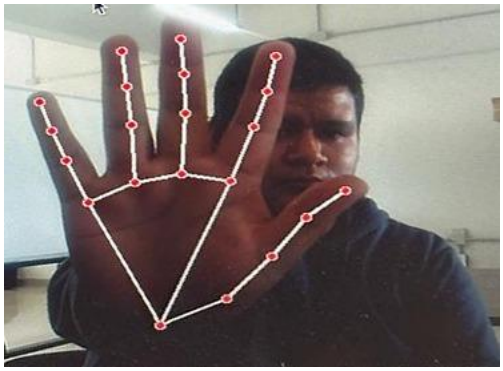

Figure 8. Gesture B grants access to the “power-on menu”.

Figure 9 shows gesture C, which initiates the shutdown menu.

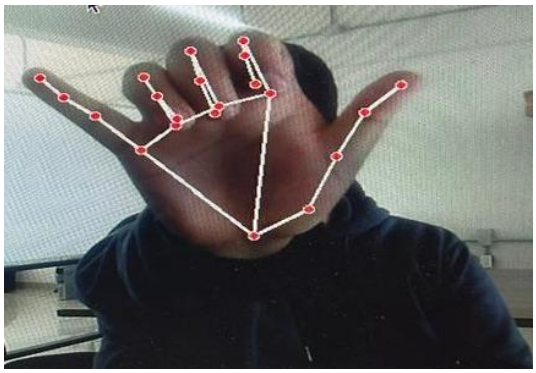

Figure 9. Gesture C, it takes you to the shutdown menu.

Figure 10 shows the gesture D, “Spotlight 1 on/off”.

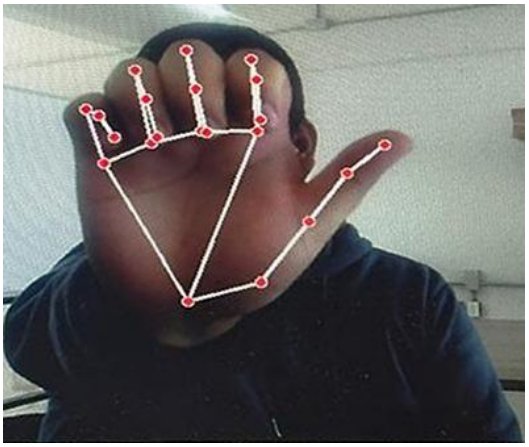

Figure 10. Gesture D, Spotlight 1 on/off.

Figure 11 shows gesture E, “Spotlight 2 on/off".

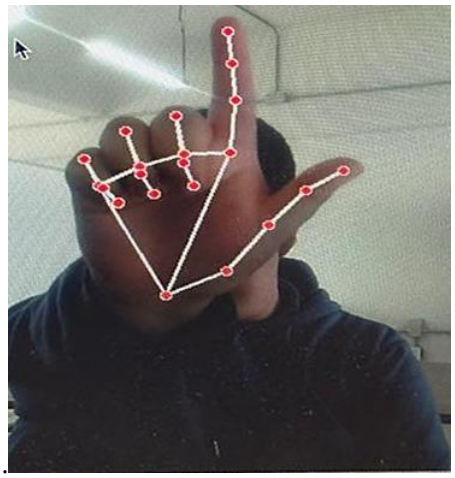

Figure 11. Gesture E, Spotlight 2 on/off.

Figure 12 shows gesture F, “Spotlight 3 on/off”.

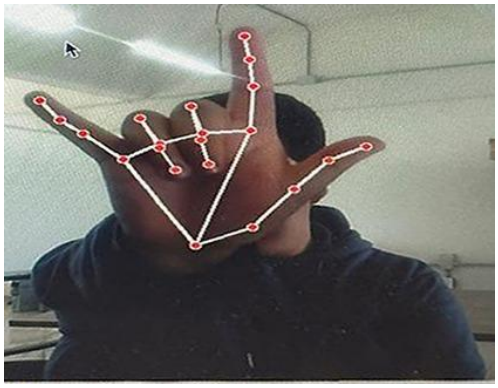

Figure 12. Gesture F, Spotlight 3 on/off.

Figure 13 shows gesture G, “Motor of blind on/off”.

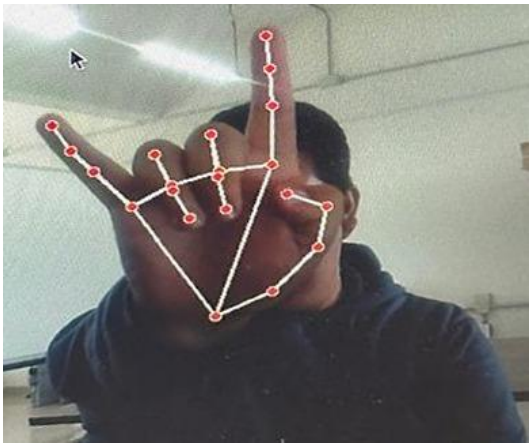

Figure 13. Gesture G, Motor of blind on/off.

Figures 14, 15, and 16 show the process of turning on light bulb 1. The sequence of gestures is a) set the system to enable state, b) enter the system of devices to turn on or off and c) turn on bulb 1.

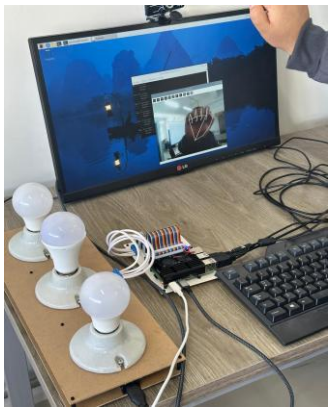

Figure 14. Gesture A, powering on the system.

Figure 15. Gesture B grants access to the “power-on menu”.

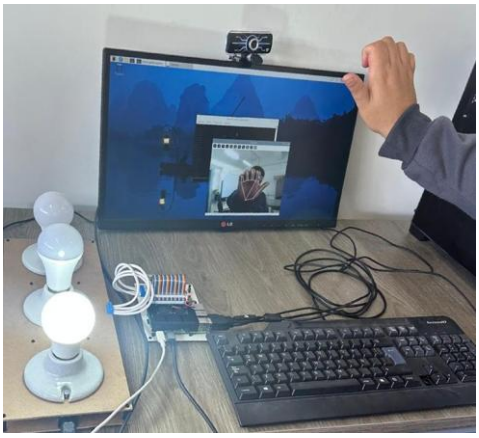

Figure 16. Gesture D, Spotlight on/off.

A. *Aspects evaluated*

Gesture recognition accuracy: It was verified that gestures were correctly recognised under different lighting conditions and viewing angles.

Response time: The time taken by the system to perform an action after detecting a gesture is almost instantaneous. The response time was fast enough to provide a smooth user experience.

Usability: The user's experience of interacting with the system was evaluated to ensure it was intuitive and did not require technical skills. It was tested with two elderly people, so further testing is needed.

Test results: A high accuracy in gesture recognition was observed during the evaluation process, being robust under varying lighting conditions. On the other hand, users found the system easy to use and they appreciated the ability to control devices without physical contact.

B. *Difference from other projects*

Focus on hand gestures: Unlike other home automation systems that rely on touch interfaces or voice commands, this project focuses on hand gesture recognition, providing a more intuitive and accessible alternative for people who have difficulty using physical devices or speaking.

Low-cost technologies: The use of computer vision technologies (OpenCV and MediaPipe) and a Raspberry_Pi 4 allows for a low-cost and highly efficient implementation, which is not common in other similar systems that typically require more expensive or complex hardware.

Optimized for different conditions: The project focuses on optimizing image processing for different lighting conditions and viewing angles, making it more robust and adaptable compared to other systems that may not perform well in variable environments.

Ease of implementation: The modularity of the system allows for easy implementation and the possibility of adding more devices in the future.

V. CONCLUSIONS AND FUTURE WORK

The implementation of the vision-controlled hand gesture recognition system demonstrated a practical use case of the Raspberry_Pi 4, combined with OpenCV and MediaPipe, to successfully develop an effective contactless control interface. The proposal shows how embedded systems can be applied to detect hand gestures in real time, and to control devices efficiently and accurately. The Raspberry_Pi 4, based on a single-chip System-on-Chip (SoC), enables features, such as processing power, the ability to run a full operating system, excellent support for Python and its libraries, and various connectivity options, making it an excellent choice for more complex and flexible embedded systems.

The use of OpenCV enables efficient image processing, while MediaPipe facilitates the detection and tracking of hand gestures, optimizing interaction with the system. This approach not only highlights the benefits of the Raspberry_Pi 4 as a computing platform, but also opens the door to future extensions in areas such as automation, accessibility and contactless control in a variety of environments, whether domestic, industrial or commercial.

The affordability of the devices used, coupled with the utilization of open-source software libraries, enables the implementation of this device as a Technology Readiness Level 6 (TR6) prototype. This approach facilitates thorough evaluation and supports the development of a commercial prototype at Technology Readiness Level 9 (TR9).

Additional feedback on system usage from the target population will be collected at a later stage to facilitate a comparative analysis with existing market solutions.

ACKNOWLEDGMENT

We thank the Centro Universitario de los Valles (CUValles) of the Universidad de Guadalajara for the space provided in its laboratories. We would also like to thank everyone who supported us to complete this research work, directly or indirectly.

REFERENCES

[1] Government of Mexico, Assistive Technologies, [Online] Available: https://www.gob.mx/profeco/es/articulos/tecnologias-asistenciales?idiom=es). [retrieved Nov. 12, 2024].

[2] V. A. Carrizal Pérez, Consumer Magazine, Digital Environment, Assistive Technologies. Jul. 21, 2021, vol. 533, ISSN 0185-8874.

[3] Global Accessibility Reporting Initiative (GARI), Learn about Mobile Accessibility: Vision. [Online]. Available: https://www.gari.info/vision.cfm?lang=es [retrieved Jul. 14, 2024].

[4] National Institute of Statistics and Geography (INEGI), Press Release No. 24/21. [Online]. Available: https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2021/EstSociodemo/ResultCenso2020_Nal.pdf [retrieved Jul. 14, 2024].

[5] National Institute of Statistics and Geography (INEGI), Population with limitations or disabilities by federal entity and type of activity performed or mental condition according to sex, 2020. [Online]. Available: https://www.inegi.org.mx/temas/discapacidad/#Tabulados [retrieved Jul. 14, 2024].

[6] World Health Organization (WHO), List of priority technical aids. Geneva, Switzerland, [Online], Available: http://apps.who.int/iris/bitstream/handle/10665/207697/WHO

_EMP_PHI_2016.01_spa.pdf;jsessionid=7643DD191B70DE7312F44E299F211411?sequence=1 [retrieved at Jul. 14, 2024].

[7] World Health Organization (WHO), Assistive Technology, [Online]. Available: https://www.who.int/es/news-room/fact-sheets/detail/assistive-technology. [retrieved Jul. 14, 2024].

[8] OpenCV. [Online]. Available: https://opencv.org/ [retrieved Jul. 14, 2024].

[9] MediaPipe Hands. [Online]. Available: https://github.com/google-ai-edge/mediapipe/blob/master/docs/solutions/hands.md. [retrieved Jul. 14, 2024].

[10] Python. [Online]. Available from: https://www.python.org/ [retrieved Jul. 14, 2024].

[11] H. Basanta, Y. -P. Huang and T. -T. Lee, "Assistive design for elderly living ambient using voice and gesture recognition system," 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff, AB, Canada, 2017, pp. 840-845, doi: 10.1109/SMC.2017.8122714

[12] J. Hossain Gourob, S. Raxit and A. Hasan, "A Robotic Hand: Controlled With Vision Based Hand Gesture Recognition System," 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), Rajshahi, Bangladesh, 2021, pp. 1-4, doi: 10.1109/ACMI53878.2021.9528192.

[13] N. Mohamed, M. B. Mustafa and N. Jomhari, "A Review of the Hand Gesture Recognition System: Current Progress and Future Directions," in *IEEE Access*, vol. 9, pp. 157422-157436, 2021, doi: 10.1109/ACCESS.2021.3129650.

[14] O. Mishra, P. Suryawanshi, Y. Singh, and S. Deokar. A Mediapipe-Based Hand Gesture Recognition Home Automation System. In 2nd International Conference on Futuristic Technologies (INCOFT), IEEE, pp. 1-6, 2023.

[15] J. Xie et al., "One-Handed Wonders: A Remote Control Method Based on Hand Gesture for Mobile Manipulator," 2024 International Conference on Advanced Robotics and Mechatronics (ICARM), Tokyo, Japan, 2024, pp. 686-691, doi: 10.1109/ICARM62033.2024.10715857.

[16] K. Singh, A. Joshi, J. Bhatt, A. Juyal and M. S. Rawat, "Webcam-Guided Gesture-Controlled Robotic Arm," 2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC), Gwalior, India, 2024, pp. 1297-1303, doi: 10.1109/AIC61668.2024.10730972.

[17] Raspberry Pi. [Online]. Available: https://www.raspberrypi.com/products/raspberry-pi-4-model-b/. [retrieved Jul. 14, 2024].

 ISBN: 978-1-68558-268-5

Ethical Risk Assessment of AI in Practice Methodology: Process-oriented Lessons Learnt from the Initial Phase of Collaborative Development with Public and Private Organisations in Norway

Natalia Murashova
Department of Teacher Education
Østfold University College
Halden, Norway
Email: natalia.murashova@hiof.no

Diana (Saplacan) Lindblom
Department of Informatics
University of Oslo
Oslo, Norway
Email: dianasa@ifi.uio.no

Aida Omerovic
SINTEF Digital
Oslo, Norway
Email: aida.omerovic@sintef.no

Heidi E. I. Dahl
Posten Bring
Oslo, Norway
Email: heidi.dahl@posten.no

Leonora Onarheim Bergsjø
Østfold AI Hub
Østfold University College
Oslo, Norway
Email: leonora.bergsjo@hiof.no

***Abstract*—Artificial Intelligence (AI) and its ethical implications are not new for academia and business. Challenges of embedding principles for ethical AI in practice are obvious and even though the gap between theory and practice is decreasing, it does not meet the urgent need for responsible technology development and deployment. Embedding ethical principles in existing risk assessment practices is a novel, process-oriented approach that can contribute to operationalising AI ethics in organisational practice. This paper elaborates on initial phase of collaborative development of ethical risk assessment of AI methodology, involving private and public organisations in Norway. We reflect upon our experience and present key takeaways in a form of three lessons learnt from embedding a Model-based security risk analysis method (CORAS) and a Story Dialog Method (SDM) in the initial phase of the collaborative methodology development. This study concludes that ethical risk assessment of AI in practice is feasible and explores design issues related to cross-sectoral settings, flexibility of the methodology, and power-relationship.**

Keywords-AI ethics in practice, ethical risk assessment, cross-sectoral collaboration, methodology development.

I. INTRODUCTION

Current transformations of the work environments due to wide introduction of AI systems call for new approaches to assess and manage benefits and risks created by these systems. AI, understood as an umbrella term in computer science for different computing techniques enabling machines to mimic "complex human skills", holds many promises and hopes together with uncertainty and fears [1][2]. Since ungoverned AI comes together with social and environmental implications, organisations that design and deploy AI are obliged to identify and mitigate possible risks throughout design process and application lifecycle.

To pursue ethical design and deployment of AI systems, a principled approach is commonly used to guide the process of development and deployment of AI systems [3][4]. Taking its popularity and relative accessibility for both practitioners and researchers, it has become a leading approach to ethical AI resulting in over 100 ethical guidelines in private and public sectors [5]-[7]. Empirical studies showed that the impact of ethical guidelines on ethical decision-making of the professionals is very low [8]-[10]. This suggests a gap between ethical AI principles available and their relevance for organisational practice. The challenge lies not only in contextualizing the principles for each stage of development of AI systems or a use case, but also in what professional competences needed to practice, promote and deploy ethical AI [11][12]. It is argued that solely ethical principles in place are not enough to responsibly navigate a complex landscape of AI applications in organisations and that embedding a "risk-oriented multi-stakeholder approach" in the assessment and management procedures is a key to effective governance of AI [13]-[16].

Prior research suggests that, to achieve ethical AI in practice, we as research community, society, governments, businesses, have to create a common language, provide equal opportunity for stakeholder involvement and create a system of incentives for ethical decision-making [6][17][18]. This study gave an opportunity to public and private organisations in Norway to take an active part in shaping ethical risk assessment of AI in practice methodology alongside with researchers from different disciplines, such as ethics, data science, risk analysis, pedagogy and social sciences. This study investigates the following research question; What are the key lessons learnt from the initial phase of ethical risk assessment methodology development? How these lessons can be applied in the next phase of the methodology development?

Through analysis of the initial stage of the ethical risk assessment of AI methodology development, this paper contributes to the developing body of knowledge on ethical AI in practice and elaborate on the outcomes of the collaborative processes involving academic, industrial and public sector partners. In addition, we reflect upon the lessons learnt to guide the methodology calibration further.

The rest of this paper is organized as follows. Section II describes the background followed by the related work section.

 ISBN: 978-1-68558-268-5

Section IV introduces CORAS and SDM as a theoretical base for this study. Section V focuses on the methods including approaches implemented during data collection, analysis and participant recruitment together with ethical consideration. Section VI presents the findings in a form of the learnt lessons. Section VII dives into discussion connected with the results together with limitations. The acknowledgement, conclusion and future work suggestions close the article.

II. Background

Rapidly emerging AI technology poses many challenges to social and organisational structures worldwide including environmental costs, social implications and ethical dilemmas [19]. Norway, among the other European countries, is on the regulatory side, adopting European Union's AI Act, but at the same time has an ambition to build a national infrastructure for AI in the public sector by 2030 [20][21]. Therefore, this invites public and private organizations to use AI systems for digitalisation and innovation [21][22].

Norwegian organisations that are developing and deploying AI systems are under pressure since they must comply with local and international data protection regulations, ethical norms, established sectorial traditions, and most importantly innovate at the same time. While legal regulations and sectoral traditions can be addressed through existing mechanisms, ethical AI is new and not well-established concept in some sectors, thus, the public and private organizations stand in front of new challenges of implementing ethical AI in practice [4][23].

In regard to the above-mentioned points, Ethical risk assessment of AI in practice (ENACT) project has a goal of creating such methodology that can benefit Norwegian businesses to evaluate and mitigate risks connected with design and deployment of AI in their organization using "ethical lens". Among the other objectives is tailoring this methodology to be adaptive, scalable, process-oriented and applicable to different organisational contexts and AI applications [24].

The ENACT consortium comprises industrial partners from public and private sectors, including medical services, finance, logistics, welfare service and education, in addition to researchers from SINTEF (lead), Østfold University College (ØUC), the University of Oslo (UiO), and the Norwegian University of Science and Technology (NTNU). ENACT is funded by the Research Council of Norway (2023-2027).

III. Related work

A. Principles for ethical AI in practice

Several governing bodies, individuals and research initiatives have released guidelines to promote responsible AI and ensure its development and deployment in an ethical manner [3][18][25]-[28]. In the last 5 years, a principled approach became the most favourable one in the literature [5][17]. Originally adopted from the bioethics and medical research, it is centred around respect of autonomy, non-maleficence, beneficence and justice [29]. Floridi has extended above-mentioned principles by adding "explicability or transparency" as a new enabling principle [25]. Policymakers, organisations, philosophers, AI researchers and practitioners have contextualized these principles resulting in over 100 public and private guidelines across the globe [5].

Despite a high level of abstraction, many attempts have been made to operationalise the ethical principles within different industrial domains [10][30][31]. Even though the organizations acknowledge importance of AI ethics and aim to be proactive, the range of employed strategies, that work in practice, seem limited, in comparison with the wide range proposed in existing literature [10][17][32]. However the gap between principle and practice is decreasing, but many organisations struggle to use ethical AI frameworks due to the ambiguity of ethical principles and variety of approaches to its design in practice [6][33].

A principled approach to AI ethics faced some criticism from being ineffective in practice to lacking a long-standing professional tradition [4][8][34]. The challenge lies not only in contextualizing the principles for each stage of development of AI systems or a use case but also in what professional competences are needed to practice, promote and deploy ethical and responsible AI [11][12].

B. Ethical risk assessment of AI in practice

Operationalising AI ethics is a challenging and multifaceted task that usually involves various stakeholder groups and processional competences. There is always an ethical aspect in a risk that is affected by our moral views and values [23][35]. Ethical Risk Assessment for identification and mitigation of possible impacts is usually performed by the ethical committees and boards and has not been standardised in regard to AI systems [36]-[38]. Due to the rapid expansion of AI use and design in organisations, the ethical dimension of risk has become a prominent topic in the discussion of social and environmental impacts of AI. Previous studies indicate that one of the challenges of ethical risk assessment of AI lies in qualitative nature of the assessment that, compared to the ordinary risk assessment approaches, is not quantifiable and rarely effective via box checking [14][36]. In addition, single sector studies showed that ethical discussions are not traditionally embedded in some sectors, compared to healthcare for example, which makes ethical risk assessment difficult and highly abstract [4][38].

Several studies have addressed the questions of how ethical risk assessment can be performed in organisations. For example, Tartaro et al. attempted to embed an ethical dimension in the risk assessment procedures through a four staged process including open ended questions, risk grouping, Likert scale evaluation (numerical value assignment), and risk identification and visualisation [14]. This study confirmed the limitations of the check-list approach and binary questions (yes/no) for complex ethical risk assessment. The study concluded that the open-ended nature of questions to prompt the participants contributes to inclusive discussion and increases the support of the risk mitigation measures by the stakeholders [14].

Felländer et al. have employed a multi-disciplinary approach to achieve data-driven risk assessment for ethical AI [39]. Using expert knowledge to build the definitions and establish the requirements for cross-sectoral application, authors highlighted the difficulty of creating ethical risk assessment tool applicable to different sectors and relatable to practical, real-world problems [39].

To address power imbalances and enhance a complex analysis of normative issues under risk assessment process, Krijger explored a relational approach by proposing a triad of decision-maker, risk-exposed and beneficiary to understand and qualitatively analyse how risk is distributed and aligned with political and social moral of the stakeholders [40].

IV. Theoretical framework

To interpret, translate and integrate ethical norms, frameworks and guidelines into organisational practice, the limitations of the principled approach can be addressed through a process-oriented perspective through using, for example, CORAS and SDM, which are described further.

A. Model-based method for security risk analysis

Compared to traditional risk analysis, which are based on "failure-oriented" aspects of a system, model-based risk assessment includes different aspects of the system, giving a holistic view on the risks connected with it [41].

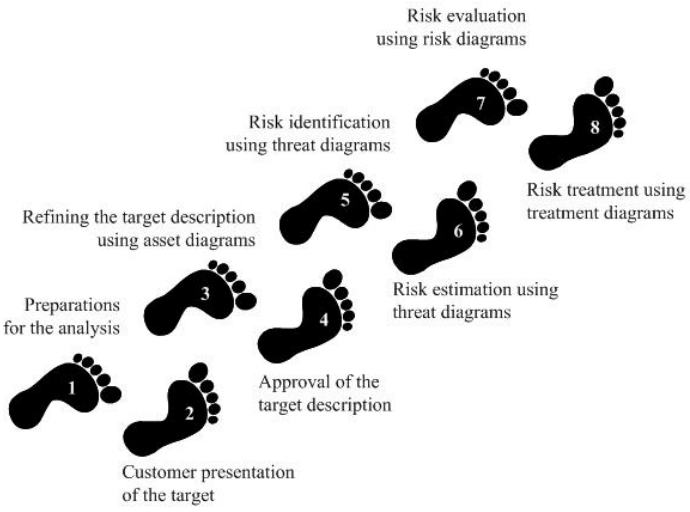

Figure 1. CORAS steps for conducting security risk analysis [42].

Model-based method for security risk assessment, also known as CORAS, is conducted in three phases: context establishment, risk assessment and risk treatment including sub-processes which contain a specific set of steps presented in Figure 1 [42].

The context establishment phase specifies the target and the scope of the analysis. During the risk assessment phase, the relevant risks are identified, evaluated and estimated. The stakeholders (organisation representatives) are being engaged on different steps of the process, allowing a variety of input and expertise shape the risk assessment outputs. The last phase focuses on treatment of the evaluated risks. Among the strengths of this approach is graphical style of the communication, visual modelling, constructive use of language and tighter integration of the assessment outputs in the system development processes [43]. As for the limitations, the practitioners noted that the CORAS language can be perceived as "too simplistic and cumbersome to use" [44].

Since this approach has been designed for defensive risk analysis of the assets with a focus on security, some fields might find it difficult to steer the discussion using such terms as "threat scenario" or "vulnerability". In addition, the risk analysis is performed by an external team which in practice is expensive and difficult to scale up.

B. Making sense of organisational experiences through Story Dialog Method

SDM is both a data collection and a data analysis method [45][46]. SDM has its roots within critical pedagogy, constructivism and feminism, and critical social sciences [45][46]. It is based on a structured dialogue and on participants' stories.

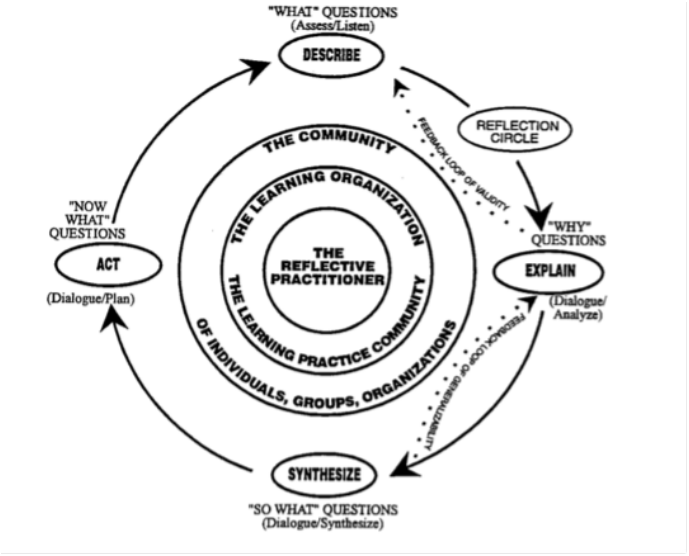

Figure 2. Story Dialogue Method [46].

A story can be described as a "self-interview" in a particular situation. Each participant has a dedicated role: storyteller, story listener, or story recorder, along with facilitators and observer roles. SDM also has values as its point of departure, which makes it relevant for the ENACT methodology. Further, the method is based on a structured dialogue, following four stages and questions related to each of the stages (see also Figure 2):

- Describe (where WHAT-type of questions are asked)
- Explain (WHY-type of questions are asked)
- Synthesis (where SO WHAT-type of questions are asked)
- Action (NOW WHAT-type of actions are asked)

While the method was initially intended to be used by a homogeneous group of participants (e.g., participants from the

same organization) in physical settings, the method lately has been used and adapted in a variety of settings with participants coming from various organizations and backgrounds, as well as online [47]-[51].

One of the benefits of this method, especially in power uneven collaborative environments, is active engagement of the participants into knowledge co-creation. It contributes to promotion of equality and inclusion in the dialog about ethical AI in organizations.

V. METHOD

In attempt to address existing limitations and challenges, ENACT methodology is collaboratively developed, tested, evaluated and validated by an interdisciplinary consortium of researchers and small and large organisations from the Norwegian private and public sector. Development of ENACT methodology is grounded in collaboration and iterative adjustment of the methods presented in the previous section.

A. Study context

The initial phase of ENACT methodology development and testing took place in September – December 2024. The working group held over 25 meetings in total (internal and with the stakeholders) to discuss the adjustments of the CORAS and integration of the SDM methods, to adopt it to digital workshop format and to accommodate the sectoral needs of the partners.

To develop a methodology for ethical risk assessment of AI which could work in practice, the working group attended to several issues. Together, the working group examined the challenges and opportunities for harmonizing and synthesising major steps of the CORAS with ethical core issues in practical settings. In addition, the analysis context was narrowed down to one use-case that all participating organisations had in common. Participating organisations have collectively agreed on using Microsoft Co-Pilot transcription tool as a use-case for ethical risk assessment.

Another important decision was made in regard to the participants and is described further.

B. Participants in a transdisciplinary collaboration

First, to cover multiple perspectives of ethical AI, the working group was comprised of the researchers with expertise in ethics, risk management, technology and pedagogy. Secondly, to attend to practical issues of business settings, the work group adopted a transdisciplinary approach. Representatives from businesses and organisations were included in the design of the ENACT methodology, and a representative from one of the businesses was included in the working group that designed the initial testing described in this paper.

Participant selection for the workshops was purposefully delegated to the ENACT-consortium business partner representatives who were our main contact persons throughout the process. These allowed us to speed up the trust bonding process with the workshop participants through the contact person who already had workplace connections with the employees [52]. To include different organisational perspectives in the collaborative process, the working group agreed on recruiting participants from different sectors and disciplines to create interdisciplinary and cross-sectoral participant pool for all three workshops.

In total, we have involved 14 unique participants in all the workshops and 29 in total across all workshops (non-unique). The workshop participants had various levels of seniority in the organisation and years of experience in their positions. Therefore, it was important to address invisible hierarchy in the group dynamic in the methodology design and try to "level up the field players" through compatible facilitation techniques and embedding SDM [53].

C. Data collection

Due to multifaceted nature of the collaborative process and the relationships that are formed, this study has comprised different data sources including: meeting notes, observational notes from the workshops, pre- and post-workshop survey, PowerPoint slides from the workshops [54]. Table I shows an overview over the data collection process.

Two facilitators and two to three observers from the working group were allocated per workshop. The facilitators were responsible for introducing the theme of the workshop, the concepts and facilitating the discussion. The observers had a passive role during the workshops but at the same time took process-oriented notes in addition to participants' reflections and ideas presented.

Together with the participants, it was decided to manually collect observational notes for securing open dialog and freedom of expression under the workshops. Several participants had made it clear before the workshop that they preferred manual data collection methods rather than audio/video recordings of the workshops, which would have affected their behavoiur. It was agreed that audio or video recording of the workshops will affect the behaviour and nature of the interactions among the participants.

Observational notes were structured according to the workshop slides (pre-selected themes) and business partner involved in the discussion (concerns, comments, ideas). In addition, we collected observers' and facilitators' post-workshop notes about the process. All the notes were gathered in a separate file after the workshops and then used for methodology refinement and adjustment of the workshop content.

D. Analysis

Using a combination of explorative approaches, such as iterative assessment of the notes, collective reflection and synthesis, helped make sense of the processes occurring in the collaborative environments [55]. Since most of the data were gathered from the participant interactions (surveys, workshops, observation notes), workshop PowerPoints and reflections of the working group, a combination of analysis techniques was applied, as follows.

TABLE I
OVERVIEW OF THE PROCESS

Workshops and CORAS steps	Workshop 1: Establishment of the context	Workshop 2: Risk assessment	Workshop 3: Risk assessment
Time and format	60 mins, Digital workshop	60 mins, Digital workshop	60 mins, Digital workshop
Structure of the workshop	1. Introduction, information and expectations 2. Key features of use case, stakeholders, timeframe 3. Values 4. High level analysis	1. Information and expectations 2. Values – identifying values and discussion 3. Scenarios – identifying values and discussion 4. Summary	1. Information and expectations 2. Scenarios 3. Group brainstorming 4. Summary of group discussion and initial measures 5. Summary and evaluation
Participants	10 participants 2 facilitators 3 observers	10 participants 2 facilitators 3 observers	9 participants 2 facilitators 2 observers
Ground for methodology adjustment	Pre-workshop survey, Work group meetings	Post workshop survey, Work group meeting, ENACT business partner meeting, ENACT project meeting	Work group meeting, ENACT business partner meeting
Concepts	Actors	Values	Scenarios
Documentation	Pre-workshop survey, 8.5 pages of structured notes	Post-workshop survey, 8 pages of structured notes	2 pages of notes

The main goal of the analysis process was to interpret textual data and researchers' observations form the workshops to explore potential lessons for ethical risk assessment of AI methodology adjustment and tuning. This study focused on a systematic process analysis that was inspired by the thematic analysis [56].

Texts that were systematically produced by the observers under the workshops and the discussions taking place right after were structured according to each workshop. These were then read multiple times by several of members of the working group, to preliminary map the process and aspects that needed adjustment. In addition, post-workshop working group meetings were held to share the experience after the workshop and reflect on the process together.

After reading the notes multiple times and summing up the reflections we have come up with analytical notes to guide the methodology adjustment and identify learning points to be taken to the next phase of testing. Sensemaking of qualitative data for this study emerged from multiple levels of analysis including participant interaction between each other, different sectors, different seniority levels. The process that we were trying to understand was unfolding in a continuum of the collaborative process rather than in a hierarchically structured manner [57]. That is why the findings are presented in a form of lessons, overarching the major take-aways from the workshops.

E. Ethical considerations

All parts of this study were conducted in line with national and international guidelines for research ethics and research integrity [58]-[60]. To attend to the rights, interests and well-being of the participants, best practices of consent, privacy and data protection were deployed. The study was planned in accordance with ENACT Data Management Plan and reported to the Norwegian Agency for Shared Services in Education and Research. All gathered data assets were stored at the Services for Sensitive Data at the University of Oslo.

In addition to protecting personal data of the participants, we had to protect organizational data too. It was therefore communicated to the participants that all the confidential information shared with us under the workshops will not be included in the overall findings and results to preserve organizational confidentiality. Moreover, the study was conducted with participants representing different organizational entities and this added another layer of complexity for ethical considerations in the research process. To create a safe environment for sharing relevant info, and to protect business representatives from sharing protected information, the researchers had several dialogues with representatives from the business partners concerning which parts of the methodology they felt safe to test in a cross-sectorial group both before and between the workshops.

VI. FINDINGS

In the course of collaborative methodology development, we have adjusted several aspects of the process to address organisational needs of the participants. The analysis of these adjustments resulted in three process-oriented lessons that are presented in this section.

A. Lesson 1. The scope of ethical risk in cross-sectoral settings

Since sectoral traditions are different, a common "analysis context" has to be identified and addressed in the methodology design to facilitate the process of risk identification and assessment. But even when the common ground is found, not all participants are ready to discuss risk treatment practices openly with other organisational sectors present. On the one hand, the difference in sector specific use-cases provided a broader scope of the discussion by incorporating different perspectives. Additionally, it helped participants to centre their reflections around ethical perspectives of risk identification and analysis and gave structure to the discussion, contributing to bonding and blending of participants' cross-sectoral experience. On the other hand, the cross-sectoral settings of the workshops created some boundaries for engagement in a deeper discussion of

the scenario-based assessment because of resistance from the participants that occurred due to different sectoral traditions and business confidentiality.

B. Lesson 2. Flexible methodology helps to address organisational needs

Our analysis suggests that ENACT methodology must accommodate organisational needs and sectoral demands in addition to addressing changing organisational dynamics and competitive AI landscape [61]. Flexibility can be achieved through feedback loops and iterative content adjustment. Feedback loops as part of the tailoring help to guide the process toward current organisational demands and make the workshops relevant for all the participants. The partner organisations wished for a methodology that can be easily embedded in everyday practice and will be resource efficient. We have observed that digital format and selected timeframe of the workshops worked satisfactory. In addition, a desired depth of the discussion was not reached despite all the adjustment efforts made by the working group. Introduction of the new concepts, used to guide the discussion, took away the time from ethical reflections and discussion.

C. Lesson 3. Easing power-relationship for structured dialog and critical reflection

Balanced and equal engagement of the participants from different sectoral traditions and seniority level was a difficult task. Our observations concluded that to promote equal power relations, the number of participants per workshop (in plenum) should be reduced to actively include everyone in the discussion. Using elements from the SDM (facilitating empowerment through giving the opportunity to all stakeholders to voice their opinions and worries) helped us to even the field players through structured yet opened environment of engagement. This approach contributed to creations of the safe space where all the participants, regardless of their position in organisation, could share their "stories" and contribute to the process. It was challenging to engage every single one of the participants at the same time, but we managed to give the opportunity to everyone to engage in the discussion and share their worries and views.

In some cases, the participants should be separated in smaller groups depending on the goal of the discussion. For fostering cross-sectoral reflections and a broader scope it is useful to separate the participant from their fellow colleagues. This fosters cross-sectoral reflections and broadens the scope, in addition to easing out existing hierarchical structures among the participants from the same organisation. On the other hand, it was admitted by the participants that they had to "hold back" some information in group and plenum discussions due to a business confidentiality.

VII. Discussion

The initial phase of collaborative development of the ENACT methodology was a demanding and rewarding process at the same time. Structural adjustments of the methodology have been enhanced by reflections of the working group and feedback loops from the participants. In the process of tuning the working group have identified following challenges that should be addressed for further development of ethical risk assessment of AI in practice including

- sectoral tradition (e.g., similarities and differences between the domain of ethics and security standards with respect to risk assessment)
- group dynamics (e.g., power dynamics in the group, business integrity, approaches to elicit organisational needs)
- confidential information of organisational practices
- format of the workshops, which had to be realistic (e.g., time, digital or physical meetings, resources required) if the businesses were to use the methodology in real world settings

When the overall workshop structure is adaptive, it can suit different professional competences, use cases and value landscapes. Two CORAS steps, including establishment of the context and risk assessment, serve as base for the ethical risk assessment and can be aligned with already established procedures of risk assessment in the organisations. Due to cross-sectoral nature of collaboration, not all participants were ready to discuss risk treatment in depth, explaining it by business confidentiality and difference in sectoral traditions. In previous studies, cross sectoral nature of the risk assessment was admitted being challenging, due to different sectoral traditions and value misalignment [38][39].

According to our results, choosing one use-case or scenario-based approach does benefit the process in term of direction to the discussion about the application of AI systems and stakeholders involved in its design and use. But at the same time, practice showed that it was difficult for facilitators to elicit concrete scenario-based solutions for ethical dilemmas that were identified in the first two steps of the assessment.

Participation and participant selection is about power, therefore the participants that are chosen to "sit at the table" matter in terms of diversity and inclusion of ethical risk assessment of AI [31]. While at this stage, we did not integrate all the SDM aspects, we adopted a few central elements from SDM: the idea of a structured dialogue, and we ensured that all participants gave their input on each question. Embedding SDM form the very beginning would, in theory, contribute to evenly distributing time for each participant to join the discussion and express their ideas during the workshop. This would even out the dynamics in uneven power environment.

In respect to the format, digital workshops worked well in terms of maximizing the number and variety of participants. As for the drawbacks, digital format created a mediated space for the interactions which had complicated the communication and the flow of the process [62]. Because of the time constrains, not all participants had the opportunity to present their ideas, in addition to that, some of the ongoing discussions had to be interrupted due to the time constrains which negatively affected depth and quality of the reflection processes.

A. Limitations

This study involved participants from a small sample of Norwegian public and private organizations. This implies that organizational culture in these sectors can differ from the international context and other local contexts, and so could the organisational needs related to AI development and deployment. Moreover, the purposeful sampling of the participants might have resulted in a biased representation of employees. In addition, the test design excluded non-Norwegian speakers and employees not involved in AI system development and use, which might be relevant stakeholders. The test also included several businesses, and to protect their interests, collaborative efforts resulted in a small sample of themes and methods agreed upon for testing.

In addition, Microsoft Co-Pilot was chosen as a common AI use-case for the cross-sectoral discussion and analysis, which limits the scope of this study to this use scenario. Observational notes taken under the workshop were taken by different researchers resulting in possible biases in the reported observations.

VIII. Conclusion and future work

Increasing presence of AI systems in everyday work of organisations signifies a need for profound changes in the way we understand, identify and assess the risks connected to its use and design. The quantity and quality of engagement with AI ethics in organisation determine the depth of the discussion and the pool of risks identified. Despite well-articulated principles for ethical AI, practice shows that their practical use in organisations is ambiguous and limited.

Our study presented several methodological reflections about ethical risk assessment of AI in practice methodology development based on Norwegian organisational context. In the course of this study, we have managed to test different workshop structures and facilitation tools based on the organisational needs that were elicited through regular feedback loops, observations and working group discussions. In a course of collaborative design, we elicited process-oriented benefits and challenges in a form of three lessons focusing on the scope of ethical risk assessment in cross-sectoral setting, flexibility of the methodology and the power relationship occurring under the workshops.

This testing phase has resulted in preliminary skeleton of ENACT methodology that needs future adjustments. Our findings suggest a need for further research on ENACT methodology implementation for example in single sector settings, different sized organisational structures and national contexts. This can enrich the understanding of the pitfalls, sector-specific value landscapes and needs that are crucial for efficient ethical risk assessment of AI in practice.

In addition, there is a need for expansion of the philosophical discourse around moral values involved in AI risk assessment and what role ethics plays in it [35].

Acknowledgment

The research project Ethical Risk Assessment of Artificial Intelligence in Practice (ENACT) is funded by the Research Council of Norway (project number 338170). The ENACT consortium comprises industrial partners from the public and private sectors, including medical services, finance, logistics, welfare service and education, in addition to researchers from SINTEF, Østfold University College (ØUC), the University of Oslo (UiO), the Norwegian University of Science and Technology (NTNU) and NORA.ai (Norwegian AI Research Consortium).

References

[1] H. Sheikh, C. Prins, and E. Schrijvers, "Artificial intelligence: definition and background," Mission AI. Research for Policy, pp. 15-41, Jan 2023, doi.org/10.1007/978-3-031-21448-6_2.

[2] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?". Proc. ACM Conference on Fairness, Accountability, and Transparency, Mar 2021, pp. 610-623, doi.org/10.1145/3442188.3445922.

[3] High-Level Expert Group on AI, *Ethics guidelines for trustworthy AI*. [Online]. Available from https://shorturl.at/2XleY 2025.04.10.

[4] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," Nature Machine Intelligence, vol. 1, no. 11, pp. 501-507, Nov 2019, doi.org/10.1038/s42256-019-0114-4.

[5] T. Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," Minds and Machines, vol. 30, no. 1, pp. 99-120, Feb 2020, doi.org/10.1007/s11023-020-09517-8.

[6] E. Prem, "From ethical AI frameworks to tools: a review of approaches," AI and Ethics, vol. 3, pp. 699–716, Feb 2023, doi.org/10.1007/s43681-023-00258-9.

[7] D. Schiff, J. Borenstein, J. Biddle, and K. Laas, "AI ethics in the public, private, and NGO sectors: A review of a global document collection," IEEE Transactions on Technology and Society, vol. 2, no. 1, pp. 31-42, Mar 2021, doi:10.1109/TTS.2021.3052127.

[8] A. McNamara, J. Smith, and E. Murphy-Hill, "Does ACM's code of ethics change ethical decision making in software development?," Proc. 26th ACM joint meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Oct 2018, pp. 729-733, doi.org/10.1145/3236024.3264833.

[9] L. Munn, "The uselessness of AI ethics," AI and Ethics, vol. 3, no. 3, pp. 869-877, Aug 2023, doi.org/10.1007/s43681-022-00209-w.

[10] J. C. Ibáñez and M. V. Olmeda, "Operationalising AI ethics: how are companies bridging the gap between practice and principles? An exploratory study," AI & Society, vol. 37, no. 4, pp. 1663-1687, Dec 2022, doi.org/10.1007/s00146-021-01267-0

[11] D. Long and B. Magerko, "What is AI literacy? Competencies and design considerations," Proc. CHI Conference on Human Factors in Computing Systems, Apr 2020, pp. 1-16, doi.org/10.1145/3313831.3376727.

[12] S. Strauß, ""Don't let me be misunderstood" Critical AI literacy for the constructive use of AI technology," Journal for Technology Assessment in Theory and Practice, vol. 30, no. 3, pp. 44-49, Dec 2021, doi.org/10.14512/tatup.30.3.44

[13] R. Clarke, "Principles and business processes for responsible AI," Computer Law & Security Review, vol. 35, no. 4, pp. 410-422, Aug 2019, doi.org/10.1016/j.clsr.2019.04.007.

[14] A. Tartaro, E. Panai, and M. Z. Cocchiaro, "AI risk assessment using ethical dimensions," AI and Ethics, pp. 1-8, Jan 2024, doi.org/10.1007/s43681-023-00401-6

[15] E. Vyhmeister, G. Castane, P.-O. Östberg, and S. Thevenin, "A responsible AI framework: pipeline contextualisation," AI and Ethics, vol. 3, no. 1, pp. 175-197, Apr 2023, doi.org/10.1007/s43681-022-00154-8.

[16] B. W. Wirtz, J. C. Weyerer, and I. Kehl, "Governance of artificial intelligence: A risk and guideline-based integrative framework," Government Information Quarterly, vol. 39, no. 4, p. 101685, Oct 2022, ISSN 0740-624X, doi.org/10.1016/j.giq.2022.101685.

[17] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices," Science and Engineering Ethics, vol. 26, no. 4, pp. 2141-2168, Aug 2020, doi.org/10.1007/s11948-019-00165-5.

[18] V. Dignum, Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Cham: Springer International Publishing AG, pp. 93-105, 2019.
[19] UN Human Rights Office of the High Commissioner, *Taxonomy of Human Rights Risks Connected to Generative AI*, [Online] Available from https://shorturl.at/KXSB4 2025.04.09.
[20] European Union, *The EU Artificial Intelligence Act*. [Online]. Available from: https://artificialintelligenceact.eu/ 2025.04.05.
[21] Norwegian Ministry of Local Government and Modernisation *National Strategy for Artificial Intelligence*. [Online]. Available from: https://shorturl.at/4oiLr 2025.04.09.
[22] K. Bjørgo. *Utnytte mulighetene i kunstig intelligens. Exploiting the opportunities of artificial intelligence*. [Online]. Available from: https://shorturl.at/Tg5ES 2025.04.09.
[23] S. O. Hansson, "How to perform an ethical risk analysis (eRA)," Risk Analysis, vol. 38, no. 9, pp. 1820-1829, Sep 2018, doi/10.1111/risa.12978
[24] ENACT. "About us". [Online]. Available from: https://www.enactai.no/about 2025.04.02.
[25] L. Floridi et al., "AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations," Minds and Machines, vol. 28, pp. 689-707, Dec 2018, doi.org/10.1007/s11023-018-9482-5.
[26] UNESCO, *Recommendation on the ethics of artificial intelligence*. [Online]. Available from: https://shorturl.at/60xL7.
[27] OECD, *Recommendation of the council on artificial intelligence* OECD/LEGAL/0449.
[28] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82-115, Jun 2020, doi.org/10.1016/j.inffus.2019.12.012.
[29] T. Beauchamp and J. Childress, "Principles of biomedical ethics: marking its fortieth anniversary (1979)," The American Journal of Bioethics, vol. 19, pp. 9-12, Oct 2019, doi.org/10.1080/15265161.2019.1665402.
[30] M. Mäntymäki, M. Minkkinen, T. Birkstedt, and M. Viljanen, "Putting AI ethics into practice: The hourglass model of organizational AI governance," ArXiv preprint, Jan 2022, arXiv:2206.00335.
[31] J. Ayling and A. Chapman, "Putting AI ethics to work: are the tools fit for purpose?," AI and Ethics, vol. 2, no. 3, pp. 405-429, Aug 2022, doi.org/10.1007/s43681-021-00084-x.
[32] B. C. Stahl, J. Antoniou, M. Ryan, K. Macnish, and T. Jiya, "Organisational responses to the ethical issues of artificial intelligence," AI & Soc, vol. 37, no. 1, pp. 23-37, Mar 2022, doi.org/10.1007/s00146-021-01148-6.
[33] J. Morley et al., "Operationalising AI ethics: barriers, enablers and next steps," AI & Soc, pp. 1-13, Feb 2023, doi.org/10.1007/s00146-021-01308-8.
[34] T. Hagendorff, "AI virtues. The missing link in putting AI ethics into practice," ArXiv preprint, Feb 2020, arXiv:2011.12750.
[35] S. O. Hansson, "Risk and ethics: Three approaches," in Arguing about Science: Routledge, 2012, pp. 629-640.
[36] A. F. Winfield and K. Winkle, "RoboTed: a case study in Ethical Risk Assessment," arXiv preprint, Sep 2020, arXiv:2007.15864.
[37] R. D. Bernabe et al., "The risk-benefit task of research ethics committees: An evaluation of current approaches and the need to incorporate decision studies methods," BMC Med. Ethics, vol. 13, pp. 1-9, Apr 2012, doi.org/10.1186/1472-6939-13-6.
[38] N. Chugh, "Risk assessment tools on trial: Lessons learned for "Ethical AI" in the criminal justice system," 2021 IEEE International Symposium on Technology and Society (ISTAS), IEEE, 2021, pp. 1-5, doi:10.1109/ISTAS52410.2021.9629143.
[39] A. Felländer, J. Rebane, S. Larsson, M. Wiggberg, and F. Heintz, "Achieving a data-driven risk assessment methodology for ethical AI," DISO, vol. 1, no. 2, p. 13, Aug 2022, doi.org/10.1007/s44206-022-00016-0.
[40] J. Krijger, "What About Justice and Power Imbalances? A Relational Approach to Ethical Risk Assessments for AI," DISO, vol. 3, no. 3, p. 56, Oct 2024,doi.org/10.1007/s44206-024-00139-6.
[41] B. A. Gran, R. Fredriksen, and A. P. J. Thunem, "An Approach for Model-Based Risk Assessment," Lecture Notes in Comput. Sci., vol 3219, 2004, doi.org/10.1007/978-3-540-30138-7_26. In International Conference on Computer Safety, Reliability, and Security, pp. 311-324. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004.
[42] M. S. Lund, B. Solhaug, and K. Stølen, Model-driven risk analysis: the CORAS approach. Springer, pp. 23-43, 2010.
[43] R. Fredriksen et al., "The CORAS framework for a model-based risk management process," Proc. 21st International Conference (SAFECOMP), Sep 2002, pp. 94-105, ISBN 978-3-540-45732-9.
[44] K. Stølen, "The CORAS method: process, concepts and notation". SINTEF. [Online]. Available from: https://shorturl.at/XTifB 2025.04.07.
[45] R. Labonté, "Reflections on stories and a story/dialogue method in health research," International Journal of Social Research Methodology, vol. 14, no. 2, pp. 153-163, Aug 2011, doi.org/10.1080/13645579.2010.492131.
[46] R. Labonte, J. Feather, and M. Hills, "A story/dialogue method for health promotion knowledge development and evaluation," Health Education Research, vol. 14, no. 1, pp. 39-50, Feb 1999, doi.org/10.1093/her/14.1.39.
[47] D. Saplacan, J. Herstad, M. N. Elsrud, and Z. Pajalic, "Reflections on using Story-Dialogue Method in a workshop with interaction design students," Proc. CEUR Workshop (AVI 18), May 2018, pp.34-43, URI https://hdl.handle.net/10642/7422.
[48] D. Saplacan, J. Herstad, A. Mørch, A. Kluge, and Z. Pajalic, "Inclusion through design and use of digital learning environments: issues, methods and stories," Proc. 10th Nordic Conference on Human-Computer Interaction, Sep 2018, pp. 956-959, doi.org/10.1145/3240167.3240264.
[49] H. W. Moen, L. E. Kjekshus, D. Saplacan, and Z. Pajalic, "How do professionals in home-based health care services react to a new trust-based way of working?–A story dialogue study," Social Sciences & Humanities Open, vol.11, p. 101251, Dec 2024, doi.org/10.1016/j.ssaho.2024.101251.
[50] D. Saplacan, T. Schulz, J. Torresen, and Z. Pajalic, "Health Professionals' Views on the Use of Social Robots with Vulnerable Users: A Scenario-Based Qualitative Study Using Story Dialogue Method," The 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN 2023), IEEE, Aug 2023, pp. 421-428.
[51] Z. Pajalic, D. Saplacan, I. Borgen, S. E. G. Olsen, and N. N. Wesseltoft-Rao, "Female university academics' reflections on the development of their academic careers in the Norwegian higher education context," Social Sciences & Humanities Open, vol. 8, no. 1, p. 100548, May 2023, doi.org/10.1016/j.ssaho.2023.100548.
[52] A. A. Kliskey et al., "Building trust, building futures: Knowledge co-production as relationship, design, and process in transdisciplinary science," Frontiers in Environmental Science, vol. 11, p. 1007105, Feb 2023, doi.org/10.3389/fenvs.2023.1007105.
[53] L. Wood and M. McAteer, "Levelling the playing fields in PAR: The intricacies of power, privilege, and participation in a university–community–school partnership," Adult Education Quarterly, vol. 67, no. 4, pp. 251-265, Nov 2017, doi.org/10.1177/0741713617706541.
[54] S. Davenport, J. Davies, and C. Grimes, "Collaborative research programmes: building trust from difference," Technovation, vol. 19, no. 1, pp. 31-40, Nov 1998, doi.org/10.1016/S0166-4972(98)00083-2.
[55] G. Guest, K. M. MacQueen, and E. E. Namey, Applied Thematic Analysis. SAGE Publications, Inc., pp. 3-20, 2012.
[56] V. Braun and V. Clarke, "Using thematic analysis in psychology," Qualitative Research in Psychology, vol. 3, no. 2, pp. 77-101, Jul 2008, doi.org/10.1191/1478088706qp063oa.
[57] A. Langley, "Strategies for theorizing from process data," Academy of Management Review, vol. 24, no. 4, pp. 691-710, Oct 1999, doi.org/10.2307/259349.
[58] E. Staksrud et al., "Guidelines for Research Ethics in the Social Sciences and the Humanities". [Online]. Available from: https://shorturl.at/inpI4 2025.04.10.
[59] The Norwegian National Committee for Research Ethics in Science and Technology, "Guidelines for research ethics in science and technology". [Online]. Available from: https://shorturl.at/k13CY 2025.04.10.
[60] P. J. Drenth, "A European code of conduct for research integrity," [Online]. Available form: https://shorturl.at/1KtpI 2025.04.10.
[61] T. H. Davenport and R. Ronanki, "Artificial intelligence for the real world," [Online]. Available from: https://shorturl.at/COinT 2025.04.10.
[62] A. Beaulieu and A. Estalella, "Rethinking research ethics for mediated settings," Information, Communication & Society, vol. 15, no. 1, pp. 23-42, Jun 2011, doi:10.1080/1369118X.2010.535838.