UR AI

Tri Rhena Tech

Alsace tech

URAI 2025

THE UPPER-RHINE ARTIFICIAL INTELLIGENCE SYMPOSIUM

AI IN MEDICINE & HEALTHCARE

FROM ALGORITHM TO CLINICAL APPLICATION ?

5 > 6 NOVEMBER

STRASBOURG (ESBS | TPS)

Interreg

Kofinanziert von der Europäischen Union
Cofinancé par l'Union Européenne

Oberrhein | Rhin Supérieur

esbs École supérieure de biotechnologie de Strasbourg Université de Strasbourg

École d'ingénieurs Télécom Physique Université de Strasbourg

ESBS & Télécom Physique
Pôle API | 300 Bd Sébastien Brant
67400 Illkirch-Graffenstaden | France

DOI https://doi.org/10.60643/urai.v2025

The Upper-Rhine Artificial Intelligence Symposium
URAI 2025
AI in Medicine & Healthcare : from algorithm to clinical application
5 & 6 November 2025
Strasbourg

Program Committee

Dominique Bagnard, *École supérieure de biotechnologie de Strasbourg*
Franz Quint, *Karlsruhe University of Applied Sciences*
Morgan Madec, *Télécom Physique Strasbourg*
Thomas Lampert, Télécom Physique Strasbourg
Ulrich Mescheder, *Furtwangen University of Applied Sciences*
Maja Temerinac-Ott, *Furtwangen University of Applied Sciences*
Enkelejda Miho, *Fachhochschule Nordwestschweiz*
Andreas Christ, *Offenburg University of Applied Sciences*
Klaus Dorer, *Offenburg University of Applied Sciences*
Henrik te Heesen, *Trier University of Applied Sciences*
Rodolfo Orjuela, *École nationale supérieure d'ingénieurs Sud-Alsace - Institut de recherche en informatique, mathématiques, automatiques et signal*
Jean-Philippe Lauffenburger, *École nationale supérieure d'ingénieurs Sud-Alsace - Institut de recherche en informatique, mathématiques, automatiques et signal*
Germain Forestier, *École nationale supérieure d'ingénieurs Sud-Alsace - Institut de recherche en informatique, mathématiques, automatiques et signal*
Alain Dieterlen, *IUT- Institut de recherche en informatique, mathématiques, automatiques et signal*
Djaffar Ould Abdeslam, *IUT- Institut de recherche en informatique, mathématiques, automatiques et signal*
Patrice Wira, *IUT- Institut de recherche en informatique, mathématiques, automatiques et signal*
Judith Kuestler, *Fachhochschule Nordwestschweiz*
Norbert Seyff, *Fachhochschule Nordwestschweiz*
Anja Huovinen, *Fachhochschule Nordwestschweiz*
Odile Lecompte, *École supérieure de biotechnologie de Strasbourg*
Bruno Kieffer, *École supérieure de biotechnologie de Strasbourg*
Volker Bucher, *Furtwangen University of Applied Sciences*
Franco Giustozzi, *Institut National des Sciences Appliquées Strasbourg*
Ali Ayadi, *IUT Robert Schuman*
Nicolas Lachiche, *IUT Robert Schuman*

Message from Dominique Bagnard, Director of ESBS

The URAI 2025 Symposium – *AI in Medicine & Healthcare: from Algorithm to Clinical Application*, held in Strasbourg, brought together researchers, clinicians, engineers and innovators to explore the transformative potential of artificial intelligence in medicine.
Co-hosted by the École Supérieure de Biotechnologie de Strasbourg (ESBS) and Télécom Physique Strasbourg (TPS), the symposium provided a valuable platform for interdisciplinary dialogue. Throughout the event, participants shared advances in research, discussed real-world applications, and reflected on the challenges involved in translating AI algorithms into clinical practice.

We were honored to welcome such a diverse and engaged international community. The quality of the discussions and the collaborations initiated during the symposium clearly demonstrated the importance of strengthening connections between life sciences, engineering and healthcare. We hope that the exchanges and ideas generated during URAI 2025 will continue to inspire new research, partnerships and innovations in AI-driven medicine.

Dominique Bagnard
Director, ESBS

Sommaire / Inhaltsverzeichnis

3D morphological reconstruction from magnetic data using deep learning

Sarah Chouchene[1], Manon Lambert[1], Corentin Féry[2], Simon Lemoigne[2], Thomas Quirin[2], Joseph Lam-Weil[1], Joris Ravaglia[1], Ralf Schumacher[3], Joris Pascal[2], Luc Hébrard[1], and Morgan Madec[1]

[1] Laboratoire ICube, UMR 7357 (University of Strasbourg / CNRS), 300 bd Sébastien Brant, F-67400 Strasbourg
chouchene@unistra.fr
[2] Institute for Medical Engineering and Medical Informatics / School of Life Sciences FHNW, Hofackerstrasse 30, CH-4132 Muttenz
[3] BellwaldTEC GmbH Muttenz, Stockertstrasse 2, CH-4132 Muttenz

Abstract. Accurate 3D reconstruction is essential for the design of cus-tomized orthotics. Traditional methods rely on expensive optical scan-ners or complex manual measurements. This work proposes an alterna-tive approach based on the use of magnetic sensors for 3D shape recon-struction. By placing magnetic sensors around the surface of the object, it becomes possible to acquire spatial information without the constraint of direct visibility. An artificial intelligence model is developed to predict and reconstruct the missing 3D point cloud, which can then be used for modeling and 3D printing orthotics. This method aims to make the man-ufacture of custom orthotics more accessible, portable, and economical.

Keywords: 3D reconstruction Magnetic sensors Deep learning Orthopedic brace 3D printing

1 Introduction

The design of customized orthoses relies on the ability to accurately reconstruct the three-dimensional geometry of the patient's limb. Traditional 3D scanning methods rely primarily on optical scanners or photogrammetry techniques, offer-ing high accuracy but requiring expensive equipment, technical expertise, and controlled lighting conditions [1]. These constraints limit their use in routine clinical settings or in resource-limited environments.

Magnetic sensors are a promising alternative to optical sensors for capturing spatial data. They enable the acquisition of positional information even in the absence of direct visibility, making it possible to take measurements in areas that are hidden or difficult to access [2]. However, the data obtained is often partial or noisy, which complicates direct 3D reconstruction.

In this context, artificial intelligence can be used to complete and reconstruct partially measured point clouds. Recent deep learning approaches applied to 3D completion exploit neural network architectures capable of modeling the spatial

structure of an object from an incomplete set of points [3]. The objective of this work is to develop an AI model capable of predicting and reconstructing the missing points in a partial cloud derived from magnetic measurements, in order to generate models that can be used for the manufacture of customized orthoses by 3D printing.

2 Methodology and results

A deep learning model is trained using supervised learning, with input from mag-netic sensors and output from optical scan references. Evaluation is performed on new shapes in order to measure the model's generalization ability. The sys-tem achieves an average error of 4.9 mm between reconstruction and reference on previously unseen shapes. However, the model does not generalize perfectly and still exhibits inconsistencies in some cases. Areas for improvement include adding parameters related to morphology and orientation, as well as combining AI with traditional rigid registration methods.

3 Conclusion

This study shows that it is possible to reconstruct 3D shapes from magnetic data using deep learning, with encouraging accuracy for a first step. The integration of additional constraints and hybrid approaches paves the way for functional and generalizable prototypes, including for shapes with complex deformations.

References

1. Haleem, A., Javaid, M.: 3d scanning applications in medical field: a literature-based review. Clinical Epidemiology and Global Health **7**(2), 199–210 (2019)
2. Ma, Y., Guo, X., Li, J., Wang, W., Chen, Y., Pan, J., Chen, J.: A survey on mag-netic sensing and communication: Technologies, sensors, and applications. IEEE Communications Surveys & Tutorials (2025)
3. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)

A Dialog-Based Multi-Agent System for Context-Aware Financial Analysis in SMEs

Miriam Velasco[1], Simone Braun[2]

[1] bimetrics GmbH
miriam.velasco@bimetrics.de [2] Offenburg University of Applied Sciences simone.braun@hs-offenburg.de

Abstract. This paper presents a dialog-based AI financial assistant designed to support small and medium-sized enterprises (SMEs) in interpreting their financial indicators. The system combines a modular multi-agent architecture with context engineering techniques (explicit agent roles, controlled function calls, and structured hand-offs) to analyze and explain financial data in natural language. A prototypical cash-flow analysis, embedded within a production platform, demonstrates how the assistant enables interactive communication of financial information, thereby improving access to managerial and financial knowledge for non-experts.

Keywords: Dialog-based AI; Multi-Agent Systems; Context Engineering; LLMs; Agentic RAG; Financial Analysis; Digital Financial Literacy; SMEs

1 Introduction

Small and medium-sized enterprises (SMEs) often lack the in-house expertise to interpret their financial metrics independently. This gap in financial literacy hinders informed decision-making and risk management [1]. Existing analytics dashboards visualize data but seldom provide accessible explanations tailored to non-experts, thereby limiting essential digital financial literacy [2].

This work introduces a dialog-based AI financial assistant that integrates large language models (LLMs) into a modular multi-agent system for grounded, natural-language analysis and explanation. Unlike generic chatbots, the system operates directly on structured financial data and provides transparent, contextualized insights. The main contribution is the design and evaluation of a scalable multi-agent architecture that combines retrieval-augmented generation, controlled function calling, and deterministic context transfer – enabling reliable, explainable financial analysis for SMEs [1-3].

2 Approach

The financial assistant is implemented as a multi-agent system integrating LLMs with targeted context engineering [3]. Specialized agents perform distinct tasks – such as intent recognition, function execution, and natural-language response generation – while maintaining a strict separation of concerns between data access, semantic processing, and output generation [4,5]. This modular design enhances scalability and reduces hallucinations by ensuring each agent operates within a well-defined context [6,7].

Figure 1 illustrates the architecture's core components, which transform structured financial information into dialog-based explanations. The system is embedded in an operational financial

platform, connected via API to real booking and document data. A cash-flow analysis was implemented as proof of concept to assess the assistant's ability to analyze and explain financial metrics interactively.

To ensure transparency and reliability in sensitive financial contexts, the system employs controlled function calling restricting LLM access to validated data from a preceding analytics layer [3]. Deterministic temperature settings (e.g., temperature = 0.3) ensure consistent tool selection, while structured handovers of contextual data between agents preserve semantic coherence across the dialogue [8]. These mechanisms form the foundation of the assistant's context-management strategy, key to reproducible, trustworthy analyses.

A user-centered evaluation (n = 10) confirmed the approach's practical applicability: all participants stated they would use the cash-flow analysis regularly, 88% rated the explanations as easy to understand, and 83% considered the results trustworthy. These findings underline that consistent context handling and transparent system logic are as important to user trust as analytical performance itself.

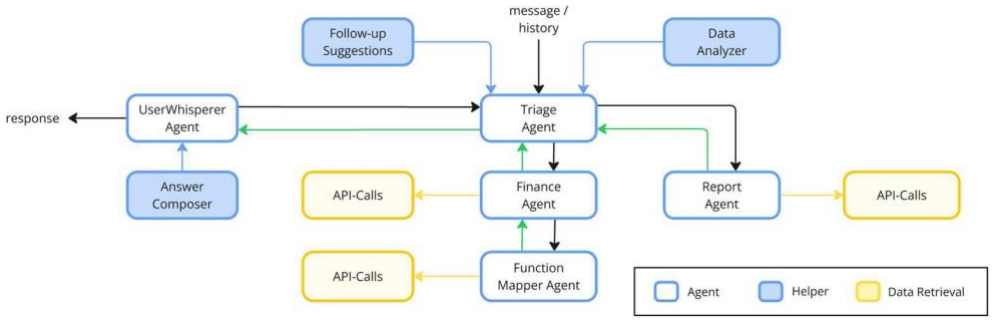

Fig. 1. Architecture of the multi-agent system underlying the financial assistant.

3 Conclusion

This paper demonstrates how combining multi-agent architectures with context-controlled LLM integration enables reliable, dialog-based financial analysis for SMEs [3-5]. The system translates structured financial data into natural-language explanations, allowing non-experts to explore business metrics interactively. The development process revealed that the architecture's effectiveness relies on clearly defined interfaces, controlled function calling, and structured context transfer between agents, which together reduce hallucinations, maintain semantic coherence, and foster user trust [3].

A user-centered evaluation confirmed that participants found the explanations highly comprehensible and the results reliable, indicating that transparent reasoning and consistent context handling are as important to user acceptance as analytical performance. The system securely integrates with real financial data and complies with privacy requirements [3-5].

While currently focused on cash-flow analysis with a small user sample, future work will extend the approach to other financial domains and a larger user base. Overall, the approach provides a reproducible blueprint for deploying LLM-based assistants in domain-specific decision-support workflows for SMEs, with significant potential to improve digital financial literacy and decision quality [1,2].

References

1. Hammer, T. (2024). Implications of Financial Literacy on Entrepreneurship in Germany. [Doctoral thesis] Universidad Católica de Murcia. URL https://repositorio.ucam.edu/bitstream/handle/10952/7984/Tesis.pdf
2. Imjai, N., Meesook, K., Somwethee, P., Usman, B., & Aujirapongpan, S. (2025). Exploring the impact of digital financial literacy to effective financial planning and control: Perspectives on competitiveness of Thai micropreneurs. *Social Sciences & Humanities Open*, *11*, 101307. doi:10.1016/j.ssaho.2025.101307
3. Mei, L., Yao, J., Ge, Y., Wang, Y., Bi, B., Cai, Y., … & Liu, S. (2025). A Survey of Context Engineering for Large Language Models. *arXiv preprint arXiv:2507.13334*.
4. Han, S., Zhang, Q., Yao, Y., Jin, W., Xu, Z., & He, C. (2024). LLM Multi-Agent Systems: Challenges and Open Problems. *arXiv preprint arXiv:2402.03578*.
5. Li, X., Wang, S., Zeng, S., Wu, Y., & Yang, Y. (2024). A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth, 1*(9), 1-43. doi:10.1007/s44336-024-00009-2
6. Lu, Y., Aleta, A., Du, C., Shi, L., & Moreno, Y. (2024). LLMs and generative agent-based models for complex systems research. *Physics of Life Reviews, 51*, 283-293.
7. Gao, C., Lan, X., Li, N., Yuan, Y., Ding, J., Zhou, Z., ... & Li, Y. (2024). Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications, 11*(1), 1-24.
8. Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019). The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Awareness and Training Offers on Digitalization and Artificial Intelligence in the German Healthcare System – an Exploratory Survey (Work in Progress)

Sophie Perret[1], Markus Schinle[2]

[1] University of Cambridge, Institute for Manufacturing (IfM), Department of Engineering
scp64@cam.ac.uk
[2] Offenburg University of Applied Sciences, Faculty of Electrical Engineering, Medical Engineering and Computer Science
markus.schinle@hs-offenburg.de

Abstract. The successful implementation of digital technologies and Artificial Intelligence (AI) in healthcare requires adequate competencies among healthcare professionals. While undergraduate medical curricula increasingly address digital health topics, little is known about awareness and training offers for healthcare staff outside of initial education. This exploratory study investigates the availability of such programs and the perceived needs for additional training in the German healthcare system. Data collection is based on an online survey, with currently 21 participants from diverse healthcare domains. Preliminary results indicate that some training activities exist, but they are often fragmented, poorly disseminated, and not tailored to specific professional needs. Respondents emphasized a strong demand for practice-oriented, interdisciplinary, and continuous training formats. As this is a work-in-progress study, further data collection and extended analysis are planned until spring 2026.

Keywords: Artificial Intelligence; Digital Health; Medical Education.

1 Introduction

The digital transformation of healthcare is progressing rapidly, with Artificial Intelligence (AI) considered a key technology to improve diagnostics, therapy, and care pathways [1]. Beyond technical and regulatory aspects, the adoption of AI in clinical practice crucially depends on healthcare professionals' competencies and acceptance. While medical curricula have started to include digital health components, the current status of awareness and training opportunities beyond formal education remains unclear. In particular, there is limited transparency regarding available programs, their uptake, and the specific needs of healthcare staff. These challenges have also been highlighted in international policy frameworks on digital health and AI [2,3].

2 Conclusion

To address the research question 'Which awareness and training offers exist regarding digitalization and AI in the German healthcare system?', an exploratory online survey was conducted. The questionnaire included closed and open-ended items on the following topics:
- Awareness and visibility of existing programs,
- Previous experience with digitalization and AI-related training,
- Perceived future needs,
- Preferred formats, duration, and thematic priorities.

At the time of writing, 21 respondents from various healthcare domains (including nursing, administration, IT, and medicine) participated. Data collection is ongoing, and results presented here are preliminary. The responses provide initial insights into the current state of awareness and training needs for digitalization and AI in German healthcare.

Respondents represented a broad age range, with the groups 25-29 years (4,8%), 30-39 years (19%), 40-49 years (19%), 50–59 years (28,6%) and 60–69 years (28,6%). Most participants identified as female (76%), followed by male (24%). The professional backgrounds were diverse, including physicians, medical students, and representatives from public health, nursing, and administrative roles. Only a minority of participants had received professional training specifically on AI: 29% reported having attended at least one training course, while 71% indicated they did not. Among those trained, the most common formats were short lectures or information sessions (<2 hours) and online modules. However, the perceived usefulness of these training courses was rated low to moderate. Respondents expressed a clear preference for interactive and flexible formats: On-site workshops (38%), Live online webinars (67%), and Self-paced online courses (57%). This indicates a demand for both digital and face-to-face opportunities, ideally in blended learning combinations. Short formats were favored: 1–2 hour sessions (62%), half-day workshops (33%), and longer-term programs such as multi-day courses or continuing education programs were also mentioned by 38% of respondents. Most participants preferred an application-oriented/practice-based level (81%), highlighting the importance of direct clinical or organizational use cases. Introductory courses were selected by 32%, while 33% indicated interest in specialized, advanced training as well as 38% in strategic management with a focus on implementation.

Thematic priorities were diverse but consistent across respondents: Large language models (81%), Ethical and legal frameworks (57%), practical implementation of AI tools in everyday work (57%), and data management and data protection in AI applications (52%). This shows a strong demand for content that combines technical insight with regulatory and ethical considerations [3,4].

Regarding responsible institutions, participants most often expected offers from: Medical associations e.g. Ärztekammer (62%), Universities and higher education institutions (38%), Professional associations e.g. Marburger Bund (29%), and Associations of Statutory Health Insurance Physicians (24%). This reflects a desire for training offers backed by trusted and authoritative institutions.

Overall, these preliminary results highlight a critical gap in the availability and visibility of structured training opportunities for AI in healthcare. The results indicate limited prior exposure to structured AI training among healthcare professionals, combined with a strong interest in accessible, practice-oriented, and interdisciplinary learning opportunities. These findings underscore the importance of coordinated, accessible training strategies to ensure the sustainable integration of AI into healthcare practice. Although isolated initiatives exist, they are often poorly disseminated and not adapted to the diverse needs of healthcare professionals. Respondents emphasized the importance of practice-oriented and interdisciplinary formats, aligning with international calls for upskilling the health workforce [1,3,5].

As the current sample is limited to 21 participants, results should be interpreted with caution. Further data collection is ongoing and expected to provide a broader and more representative picture by spring 2026. Future analyses will explore subgroup differences (e.g., between clinical and administrative staff) and examine correlations between digital competence levels and training needs.

References

1. WHO (2021). Ethics and governance of artificial intelligence for health. ISBN: 978-92-4-002920-0.
2. European Commission (2023). Digital Health and Care. https://health.ec.europa.eu/ehealth-digital-health-and-care/digital-health-and-care_en, last accessed 19.09.2025.
3. Longhini, J., Rossettini, G., & Palese, A. (2022). Digital Health Competencies Among Health Care Professionals: Systematic Review. Journal of Medical Internet Research, 24(8), e36414. doi:10.2196/36414
4. Alotaibi, N., Wilson, C. B., & Traynor, M. (2025). Enhancing digital readiness and capability in healthcare: a systematic review of interventions, barriers, and facilitators. BMC Health Services Research, 25, Article 500. doi:10.1186/s12913-025-12663-3
5. Mainz, A., Hess, R., & Pohlmann, S. (2024). Measuring the Digital Competence of Health Professionals: A Scoping Review. JMIR Medical Education, 10(1), e55737. doi:10.2196/55737

Beyond Single Models: Unsupervised Ensemble Selection for Small Language Models in Medical QA

Nicolas Ventulett, Fabian Nicklas, Eric Gaida, Dieter Wallach, and Jan Conrad

University of Applied Sciences Kaiserslautern
{nicolas.ventulett, fabian.nicklas, eric.gaida, dieter.wallach, jan.conrad}@hs-kl.de

Abstract. Small Language Models (SLMs) provide efficient alternatives to large models for clinical open-ended question answering (QA) but often show variable performance. We propose two unsupervised answer selection strategies for SLM en-sembles: a confidence-based method using normalized perplexity and a consensus-based medoid method capturing semantic similarity among model outputs. Eval-uations on three clinical QA benchmarks show that both strategies outperform single-model and random selection baselines. The results show that unsupervised confidence and consensus mechanisms can enhance the performance of SLM en-sembles for medical QA without requiring additional training or increasing model size.

Keywords: Artificial Intelligence, Large Language Models, Small Language Mod-els, Medical Question Answering, Clinical NLP

1 Introduction

Recent advances in Large Language Models (LLMs) have shown impressive performance in question answering (QA) on clinical documents, enabling patients and clinicians to ef-ficiently extract information from medical records. However, deploying large-scale LLMs in healthcare scenarios remains challenging due to strict privacy regulations and com-putational cost. Consequently, small language models (SLMs) are gaining attention for on-premise or on-edge deployment. While SLMs tend to show greater variability and lower accuracy compared to LLMs, ensemble methods that aggregate the predictions of multiple SLMs offer a promising way to improve performance without increasing model size. This work presents two unsupervised answer selection strategies for SLM ensem-bles in medical QA: a confidence-based method using normalized model perplexity and a consensus-based medoid method capturing semantic similarity among model outputs. Both approaches are evaluated against random selection in ensembles and single-model answering as baselines.

2 Related Work

Ensemble methods have demonstrated improvements in language model performance over various tasks, however most approaches focus on large-scale models or require su-pervised training [1][2]. In the field of medical QA, recent work has achieved expert-level performance through domain-specific fine-tuning and ensemble refinement [3][4]. While ensemble approaches show promising improvement in model performance, unsupervised output selection strategies for SLM ensembles in a clinical context for open-ended QA remain largely unexplored.

3 Methodology

The ensemble consists of six state-of-the-art instruction-tuned SLMs with ≤ 7B param-eters: SmolLM3-3B [5], Qwen3-4B-Instruct [6], Phi-4-mini-instruct (4B) [7], DeepSeek-R1-5B [8], Gemma-3-4B-it [9] and Mistral-7B-Instruct-v0.3 [10]. Each model in the en-semble receives the same structured prompt including a patient discharge summary and a corresponding question to be answered based on the clinical note, resulting in six can-didate answers to select from. We then compare two unsupervised ensemble selection strategies: The first strategy selects the candidate with the highest model confidence, estimated through perplexity scores computed during generation. Since raw perplexity is not comparable across different model architectures, we employ bits-per-byte (BPB) as byte-normalized perplexity variant [11]. The answer with the lowest BPB value is selected as the most confident prediction and final output. The second strategy selects the answer that is semantically most similar to all other ensemble answers. For each candidate, we compute its sentence embedding [12] and measure its average cosine similarity to all other candidates in the ensemble. The answer with the highest average similarity is selected as the final output. This consensus-based strategy selects the most representative (medoid) answer from the ensemble, under the assumption that agreement among diverse models indicates higher reliability.

4 Evaluation

The proposed methods were evaluated on three different benchmark datasets: EHR-DS-QA (506 samples) [13], MeDiSumQA (416 samples) [14] and CliniQG4QA (1287 samples) [15]. These datasets contain human-verified and medically relevant QA-pairs on deiden-tified discharge summaries from MIMIC-III/-IV clinical notes [16][17]. All SLMs and ensemble methods were evaluated using BERTScore F1, capturing semantic similarity between prediction and ground-truth answer [18]. We compared against two baselines: random ensemble selection and individual model performance without ensembling. Across all three datasets, both ensemble methods consistently outperformed random answer se-lection and single-model inference across all six SLMs. The semantic medoid approach achieved higher mean BERTScore F1 across all samples than any individual model, while the BPB-based confidence method showed the best overall performance across all benchmarks. On the EHR-DS-QA dataset, for example, the confidence-based selection improved mean BERTScore F1 by 7.3 points (0.8441 vs. 0.7674) compared to DeepSeek-R1-5B single-model answering.

5 Conclusion

This work demonstrates that unsupervised ensemble selection strategies, confidence-based and consensus-based, can effectively enhance the performance of SLMs in clini-cal open-ended question answering, consistently outperforming single-model and random selection baselines across three benchmark datasets. The results highlight a practical approach for privacy-sensitive healthcare scenarios that require on-premise or on-edge deployment while maintaining a balance between computational efficiency and answer quality. Future research should investigate adaptive or dynamic ensemble selection meth-ods to further reduce computational overhead.

References

1. Ashiga, M., Jie, W., Wu, F., Voskanyan, V., Dinmohammadi, F., Brookes, P., Gong, J., Wang, Z.: Ensemble learning for large language models in text and code generation: A survey (2025)
2. Chen, Z., Cano, A.H., Romanou, A., Bonnet, A., Matoba, K., et al.: Meditron-70b: Scaling medical pretraining for large language models (2023)
3. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., et al.: Toward expert-level medical question answering with large language models. Nature Medicine **31**(3) (March 2025) 943–950
4. Lucas, M.M., Yang, J., Pomeroy, J.K., Yang, C.C.: Reasoning with large language models for medical question answering. Journal of the American Medical Informatics Association **31**(9) (September 2024) 1964–1975
5. Bakouch, E., Ben Allal, L., Lozhkov, A., Tazi, N., Tunstall, L., et al.: SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3 (2025)
6. Team, Q.: Qwen3 technical report (2025)
7. Microsoft, :, Abouelenin, A., Ashfaq, A., Atkinson, A., Awadalla, H., et al.: Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras (2025)
8. DeepSeek-AI: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (2025)
9. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., et al.: Gemma 3 technical report (2025)
10. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., et al.: Mistral 7b (2023)
11. Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., et al.: The pile: An 800gb dataset of diverse text for language modeling (2020)
12. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Pro-cessing, Association for Computational Linguistics (11 2019)
13. Kotschenreuther, K.: EHR-DS-QA: A Synthetic QA Dataset Derived from Medical Dis-charge Summaries for Enhanced Medical Information Retrieval Systems (2024)
14. Dada, A., Koras, O.A., Bauer, M., Butler, A., Smith, K.E., Kleesiek, J., Friedrich, J.: MeDiS-umQA: Patient-Oriented Question-Answer Generation from Discharge Letters (February 2025) arXiv:2502.03298 [cs].
15. Yue, X., Zhang, X.F., Yao, Z., Lin, S., Sun, H.: CliniQG4QA: Generating Diverse Questions for Domain Adaptation of Clinical Question Answering (December 2021) arXiv:2010.16021 [cs].
16. Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L.A., Mark, R.: MIMIC-IV (2024)
17. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., et al.: Phys-ioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation **101**(23) (2000) e215–e220 Publisher: Am Heart Assoc.
18. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert (2020)

Clinically-Ready Label-Flip Detection for Medical AI

Daniel Schönle [1], Christoph Reich [2]

[1] Affiliation
daniel.schoenle@hs-furtwangen.de
[2] Affiliation
christoph.reich@hs-furtwangen.de

Abstract. Medical AI pipelines face integrity risks from label flipping—mislabeling that harms thresholds, calibration, and parity. Because anomalies are rare, evolving, and often mislabeled, a purely supervised detector tends to miss new problems and flood reviewers with false alarms; a triage loop—rank strong model-vs-label disagreements, review a small top slice, fix, retrain—keeps effort low and results trustworthy. We present a lightweight procedure: basic plausibility/duplicate checks; leakage-safe K-fold cross-fitting; calibration; and Confident Learning to derive per-example flip scores (and the confident joint). High-scoring cases receive budgeted chart-review; we then selectively relabel or reweight, retrain, and recalibrate. We evaluate flip-ranking (PR-AUC, precision@k, TPR@low-FPR) and downstream AUROC/PR-AUC, ECE/Brier, and parity deltas. A HiRID ICU case demonstrates integrity and calibration gains with limited review effort.

Keywords: : label noise; confident learning; calibration; medical AI; fairness; governance; anomaly detection

1 Introduction

We target label flipping in EHR datasets used for model development & validation. Adversaries or process faults can inject or flip labels at data entry, curation, or during merges. Security goals are to maintain discrimination, calibration, & subgroup parity, & to provide auditable evidence trails (provenance, adjudication notes, versioned lineage). EHR-specific flip vectors include ICD miscodes & mapping drift, rule-based phenotype errors, timestamp/encounter misalignment, & contradictions between billing codes & clinical evidence (vitals/labs/notes).

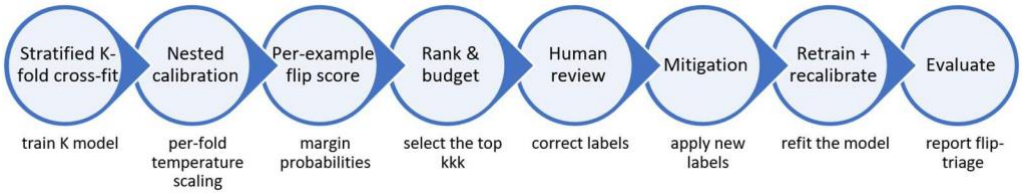

Figure 1. Eight-step ML-only triage loop for label-flip detection.

Blueprint: (1) Ingress rules check provenance, duplicates, mutual exclusivity, value ranges, unit harmonization (e.g., mg/dL↔mmol/L), physiologic plausibility, & encounter/windowing consistency; (2) Disagreement analytics from targeted chart-review & code–evidence conflict checks; (3) Leakage-safe K-fold cross-fit to obtain out-of-fold probabilities; (4) Calibration on a nested holdout (temperature by default; Dirichlet optional under imbalance); (5) Confident Learning to compute the confident joint (CJ) & estimate T; (6) Rank suspected flips where label & calibrated predictions disagree w/ high confidence; (7) Mitigate via budgeted relabeling,

reweighting, or relabel/weight w/ T; (8) Retrain, recalibrate, & log decisions. We use gradient-boosted trees (LightGBM/XGBoost) on benchmark features.

2 Evaluation Protocol

Flip ranking: PR-AUC, ROC-AUC, TPR@1% FPR, precision@k under review budgets, & lift over random; report bootstrap confidence intervals. Downstream (before/after mitigation): AUROC/PR-AUC, ECE & Brier, reliability diagrams, & subgroup parity deltas (e.g., ΔFNR). Subgroup analyses use ICU-relevant strata (sex, age bands, unit type, SOFA/APACHE bins, admission year) w/ temporal splits to approximate deployment drift; top-k suspects can be chart-reviewed against notes/labs when available.

3 Datasets & Experiment

Detectors only triage candidates; labels change only after expert adjudication; the system does not make clinical decisions. HiRID
HiRID access requires PhysioNet credentialing & a data-use agreement; data are de-identified. Limitations include site/process shifts across years/units & potential bias in Tunder covariate shift; subgroup-specific T can reduce bias at the cost of variance.

Primary dataset: HiRID (high-resolution ICU EHR; PhysioNet). We use the published HiRID-ICU-Benchmark pipeline to materialize features & tasks, minimizing engineering effort. Tasks: (i) Flip detection on a noisy, code-derived phenotype (e.g., pneumonia/CHF) constructed w/ the benchmark's labeling rules; (ii) Downstream evaluation on a stable outcome (in-hospital mortality or intervention onset). Pipeline: benchmark features → LightGBM/XGBoost w/ 5-fold cross-fit → temp-scaling (Dirichlet optional) → Confident Learning (CJ, T) → budgeted relabeling (top 1–2%) or relabel/weight → retrain + recalibrate.
Replication dataset: eICU Collaborative Research Database (eICU-CRD) can be processed w/ the same pipeline to validate cross-site robustness.

References

1. Northcutt, C. G., Jiang, L., & Chuang, I. (2021). Confident Learning: Estimating uncertainty in dataset labels. Journal of Artificial Intelligence Research.
2. Patrini, L., Rozza, A., Menon, A., Nock, R., & Qu, L. (2017). Making deep neural networks robust to label noise: A loss correction approach. CVPR.
3. Hyland, S. L., Faltys, M., Hüsser, M., et al. (2020). HiRID: A high-time-resolution ICU dataset for research. Scientific Data.
4. Goldberger, A. L., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet. Circulation.
5. HiRID-ICU-Benchmark (GitHub repository and accompanying paper): Reproducible tasks and pipelines for HiRID.

Cognitive Load Estimation through Eye-Tracking in Industrial Tasks

Kanan Gurbanov[1], Cedric Bobenrieth[1,2], Nathalie AlMakdessi[1,2], Farid Kacimi[1,2], Grégoire Chabrol[1,2], Samy Rima[3], Rabih Amhaz[1,2]

[1] ICube Laboratory, UMR 7357, University of Strasbourg, France
amhaz@unistra.fr
[2] Icam site de Strasbourg-Europe, France
name.surname@icam.fr
[3] Schmid Research Group, University of Fribourg, Switzerland
samy.rima@unifr.ch

Abstract. Understanding and managing cognitive workload is critical for safety and efficiency in industrial settings, yet practical measurement techniques are limited on the factory floor. We present a vision-based approach to estimate cognitive load using eye-tracking, hand tracking, and object detection, avoiding intrusive sensors like EEG. We evaluated this method in a real-world assembly task performed by an expert under two conditions: an organized workspace and a disorganized workspace. A mobile eye-tracker recorded the worker's gaze, while computer vision detected hands and tools in use. The disorganized condition elicited higher visual workload, evidenced by more frequent fixations and saccades, broader gaze dispersion, and more attention to irrelevant areas, despite little change in physiological proxies such as pupil size and blink rate. These results demonstrate that our non-intrusive, vision-only system can distinguish cognitive workload differences in an industrial task, laying the groundwork for in-situ workload monitoring without requiring cumbersome biosensors.

Keywords: Cognitive Workload, Eye-Tracking, Object Detection, Industrial Ergonomics, Human-Machine Interaction.

1 Introduction

Cognitive overload occurs when an operator's mental resources are exceeded by task demands, often leading to errors or slowdowns. In industrial environments, human performance under high cognitive load is a major factor in safety and productivity. Traditional measures (e.g., EEG-based monitoring) are impractical for real-time use on a production floor. There is a growing need for lightweight, real-time workload measures that can be seamlessly integrated into industrial workflows.

Many existing frameworks for workload detection rely on multimodal biosignals (EEG combined with eye-tracking or other sensors) and complex lab setups. Moreover, prior models often do not target the predominantly visual-perceptual nature of tasks like assembly and inspection. In such tasks, an operator's cognitive load may stem largely from visual search and attention management, rather than abstract reasoning alone. This work explores whether reliable indicators of cognitive workload can be obtained from vision-based data alone, specifically using an operator's eye movements, hand interactions, and observed tool usage, without the need for EEG or other invasive sensors.

We developed a proof-of-concept approach combining mobile eye-tracking with computer vision to monitor a worker's visual attention and actions during an assembly task. An expert technician

performed the same cart assembly task under organized and disorganized conditions to isolate the effect of workspace layout on cognitive load.

We hypothesized that the unorganized workspace would impose higher cognitive load, observable as more extensive visual search behavior (e.g. more fixations, wider gaze spread, longer task time), whereas the organized workspace would allow more efficient, focused visual attention.
In this paper, we report how eye-tracking metrics and interaction data differed between these conditions, and what they reveal about cognitive workload. To our knowledge, this is one of the first demonstrations of using eye-tracking with scene object detection to assess cognitive load in an in-situ industrial assembly task. The results show that subtle changes in workspace layout can be quantified through vision-based measures. Our work contributes an integrated analysis pipeline for multimodal workload monitoring and empirical evidence that a non-intrusive, video-based approach can capture cognitive load differences in real manufacturing conditions.

2 Related Work

This section reviews relevant literature in three domains: (1) eye movement as an indicator of cognitive load, (2) multimodal and physiological workload measurement, and (3) vision-based monitoring in industrial settings. We summarize prior achievements and identify the remaining gaps that motivate our study.

Eye Movement as a Proxy for Cognitive Load

Eye movements (fixations, saccades, pupil dilation, microsaccades) are widely used in cognitive and HCI studies to infer mental processing (e.g. decision-making, memory retrieval), as gaze is tightly linked to attention and information acquisition. A classic study by Krejtz et al. (2018) compared pupil dilation and microsaccade metrics in arithmetic tasks, finding that both signals discriminated task difficulty, though each had distinct sensitivity profiles. [1] Many studies assume a monotonic increase in pupil diameter with load (the so-called task-evoked pupillary response). More recently, some works differentiate intrinsic vs. extraneous cognitive load using oculometric signals. For instance, a 2025 study used eye-tracking, heart-rate variability, and galvanic skin response to classify intrinsic vs. extraneous load in multimedia tasks, achieving promising predictive power. [7] This is relevant: our manipulation (workspace clutter) is akin to extraneous load, and we show gaze metrics reflect it (even when physiological signals don't shift strongly). In serious game contexts, other researchers have grounded measurement in theory, calibrating eye metrics to time-based resource-sharing models of load. E.g. a study with 42 participants playing time-critical resource-management games mapped attentional demand to gaze statistics. [6] Such work shows the feasibility of interpreting gaze within formal models of load, but still typically in constrained, quasi-lab settings rather than real-world manufacturing. Thus, while ocular metrics are well explored, there are few studies that apply them in real industrial task contexts, especially combining gaze with scene understanding (object interactions, hand movements) to localize sources of load.

Multimodal and Physiological Workload Measurement

To capture cognitive load robustly, many works fuse multiple sensors. For example, eye tracking combined with EEG or ECG yields higher classification accuracy in cognitive load estimation than

any single modality alone. One survey notes fusion of eye, EEG, and GSR is common in ergonomics and usability studies [8]. A recent system-level work, CLERA (2023) in [2], proposes a unified deep model for joint eye-region analysis and cognitive load estimation “in the wild” (i.e. less constrained settings). CLERA jointly learns eye landmarks, blink prediction, pupil estimation, and cognitive load, and shows improved robustness over isolated pipelines. However, CLERA is built for HCI-style tasks, not industrial hand–tool workflows. It does not explicitly model object interactions or hand movement in the scene, which are crucial in an assembly task. Another approach, in the VR/training domain [9], fuses gaze with heart rate variability (HRV) and other signals to detect both cognitive load and stress, enabling adaptive systems. These systems are promising, but again not tailored for noisy, real-world factory conditions with dynamic tool usage and occlusions. Thus, while multimodal methods improve robustness, they often sacrifice deployability. Our approach intentionally remains vision-only, trading off some signal richness for practicality.

Vision-Based Monitoring in Industrial Environments

A separate but related branch of research uses computer vision to monitor worker behavior, safety compliance, and ergonomics. For instance, in construction, vision systems detect helmet usage, unsafe postures, or proximity violations [10]. A 2024 work introduces a vision-based framework for human behavior monitoring in a car door assembly line, combining multi-camera video and 3D motion capture to build a dataset (CarDA) for analyzing assembly actions. [4] That system, however, focuses on what actions are performed (pose, object presence), not why (i.e. underlying cognitive load). Other recent vision-based industrial systems target ergonomic risk assessment using skeleton estimation and motion features to flag musculoskeletal risk. For example, Agostinelli et al. (2024) propose semi-automated ergonomic risk assessment in manufacturing using depth sensors and computer vision. [11] These systems are orthogonal to cognitive load measures, they do not access gaze or attention. A recent review on computer-vision-based biomechanical and workload assessment highlights that many methods infer physical or biomechanical load (e.g. motion energy, joint torque proxies) but rarely estimate cognitive or perceptual load from video alone [3]. Furthermore, these reviews note that fusing vision with physiological or behavioral signals remains an underexplored frontier.

We can conclude from related work that most eye-tracking research remains in lab or UI settings; few studies embed gaze in complex physical tasks. Multimodal methods are powerful but often less deployable, especially outside controlled environments. Vision-only systems in industry focus on motion or safety, not cognitive load or attention dynamics. No prior work (to our knowledge) fuses gaze, hand movement, and object detection in a real industrial assembly task to estimate cognitive load, especially with controlled manipulations of environmental clutter. Our study tries to fill this gap: by combining eye-tracking with scene-level understanding (hands, tools) in a real task, we demonstrate that gaze metrics can reliably detect visual workload changes in situ, bridging the sensor-rich lab world and the vision-only industrial world.

3 Method

Participants & Task. We recruited one expert assembly technician, Robert Bolusset (LEAN manufacturing consultant and trainee), to perform a cart assembly task under two workspace conditions: organized (tools neatly laid out) and unorganized (same tools, scattered randomly) cf.

Fig 1. The task procedure remained identical in both trials, ensuring any behavioral differences stem from layout effects.

Fig. 1. Example frames from organized vs unorganized videos

Apparatus & Data Collection. The participant wore a mobile Pupil Labs eye tracker, capturing synchronized scene video and gaze overlay streams (fixations derived from raw gaze). We processed video frames with MediaPipe for hand detection and YOLOv8 (trained on 850+ annotated images across 11 object classes) for tool/part detection (achieving ~87% mAP at the validation threshold). We synchronized gaze, fixation ID, object presence, and hand interaction per frame into a structured JSON timeline dataset.

Fig. 2. Our expert during the assembly process wearing pupil lab neon Eye-tracker.

Measures & Analysis. From this timeline, we extracted:

- **Eye metrics:** total fixations, average fixation duration, saccade count, gaze dispersion, pupil diameter, blink rate
- **Attention allocation:** proportion of fixations on relevant vs. irrelevant areas, AOI (area of interest) transition frequency, dwell times on key objects
- **Task performance:** total completion time, any corrections or tool mis-selections

We compared these metrics between organized and unorganized trials in a within-subject, descriptive analysis.

4 Results

The disorganized workspace produced clear, consistent divergences in gaze behavior and attention distribution relative to the organized layout:

- **Fixation & saccade counts** increased by ~24%, from 494 to 613 fixations, and 493 to 612 saccades, respectively, indicative of a more fragmented scanning strategy.
- **Average fixation duration** dropped slightly from ~563 ms (organized) to ~550 ms (unorganized), reflecting shorter glimpses and more frequent transitions.
- **Gaze dispersion** expanded from a radius of ~158 px to ~190 px, confirming broader spatial search.
- **Attention misallocation:** The share of fixations landing on irrelevant areas rose from 62.0% to 73.6%, confirming that clutter forces more "wasted" visual glances.
- **AOI transition frequency** also rose, signaling more gaze switching between items and distractors.
- **Pupil diameter** and **blink rate** remained effectively unchanged (~5.23 mm, ~11 blinks/min), suggesting that while visual scanning load increased, overall arousal or stress did not escalate detectably for this expert.
- **Completion time** increased in the disorganized condition (fitting with the increased visual workload), though no severe errors or tool misuses occurred, consistent with the participant's expertise.

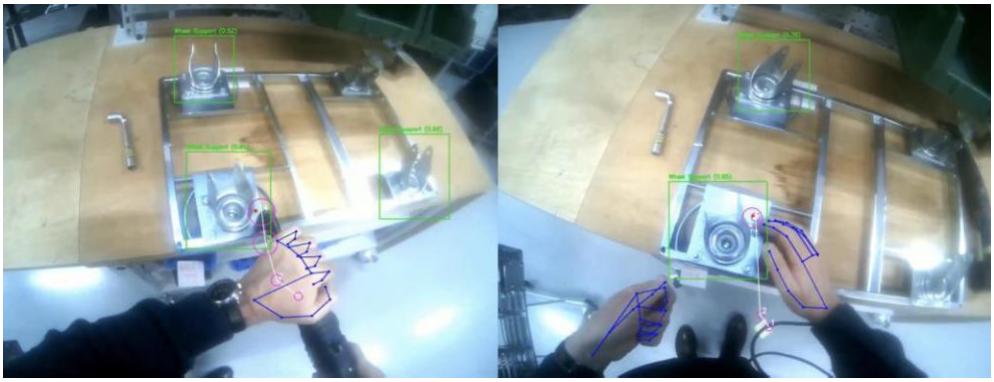

Fig. 3. Result video combining data between industrial object detection, hand tracking, and eye fixation and transition of the eyes (purple circles, size of the circles is the duration of the fixation)

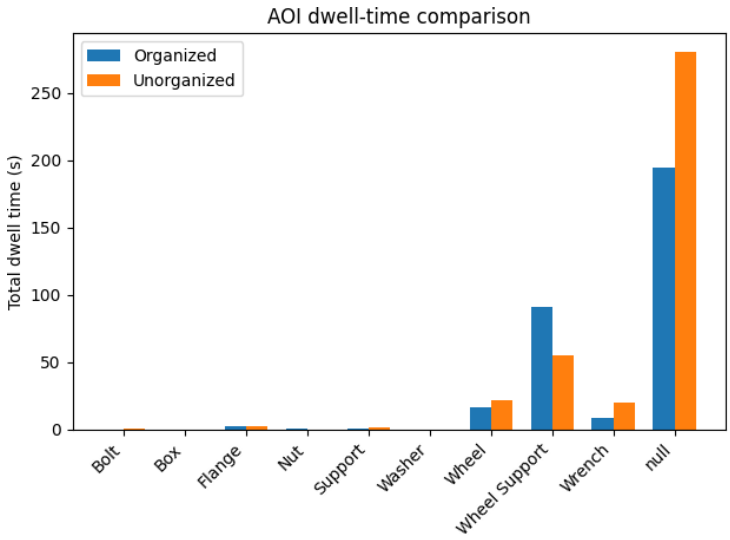

Fig. 4. AOI dwell-time comparison between organized and unorganized workplace correlated with industrial object detection

These differences align with theoretical expectations: increasing extraneous visual demands (via clutter) forces more scanning and attentional shifts, which we were able to detect through gaze metrics alone. The absence of strong changes in pupil/blink measures implies that for an expert operator, the extra visual load remained manageable in terms of overall cognitive stress.

5 Discussion

Our results validate that workspace layout directly influences visual/perceptual load, observable through eye metrics. In cognitive load terms, the disorganized layout introduced extraneous load, effort invested in navigating the environment rather than intrinsic task logic. The increase in fixations, saccades, gaze dispersion, and AOI switching signals this overhead.

Interestingly, the stable physiological metrics (pupil, blink) suggest that the mental load ceiling was not breached. Because our participant was an expert, they likely compensated for extra visual demands without triggering elevated arousal. This aligns with research showing that experts often absorb increased perceptual demands via compensatory strategies before cognitive overload manifests (e.g. in pupil dilation) in novices.

Relative to prior work, Krejtz et al. demonstrated that pupil and microsaccade signals track mental difficulty, but they do so in controlled tasks, not physical tasks with environmental clutter. CLERA offers an elegant vision-only model for cognitive load in “the wild,” but lacks a coupling to object interaction or hands, something essential in assembly tasks. Works in industrial vision largely focus on action or safety detection rather than the why (i.e. internal cognitive strain). For example, CarDA monitors worker actions but not attention or cognitive load.

Thus, our work treads a new space: vision-enabled cognitive load estimation in a manual industrial context. We show that eye–scene fusion yields signals of perceptual workload in a realistic environment.

Limitations:

- Single expert subject: Results reflect conditions for a highly skilled user; novices might show different coupling between visual and physiological signals.
- Task specificity: We tested one assembly scenario (cart). Generalizing to other tasks (e.g., inspection, wiring) requires further validation.
- Lighting and tracking noise: Real workshop conditions can degrade gaze or object detection; occasional frame drops or misalignments may introduce noise.

Implications & Insights

Deployment feasibility: Our approach (eye + video) is far less intrusive than EEG, making it a more viable candidate for industrial deployment.
Explainability of load signals: Since we link gaze to objects and hands, we can interpret where the extra load originates (e.g. scanning distractors), not just infer a black-box workload score.
Potential for adaptive feedback: Even without live feedback in this study, the pipeline could support continuous monitoring and alerts when visual load becomes excessive.

6 Conclusion

We have demonstrated that a vision-based system, combining mobile eye-tracking, hand detection, and object recognition, can detect meaningful differences in visual workload arising from workspace organization in an industrial task. Despite the same underlying assembly procedure, the cluttered layout imposed higher perceptual demand, evident via gaze metrics (fixation count, dispersion, AOI distribution). Crucially, this was achieved without physiological sensors.

Our findings suggest that vision-only workload sensing is viable in pragmatic settings, especially for tasks where visual search and attention dominate. The next steps include validating across more participants and tasks, calibrating thresholds to individual operators, and exploring real-time deployment.

References

1. Krejtz, K., Duchowski, A. T., Niedzielska, A., Biele, C., & Krejtz, I. “Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze.” PLoS ONE, 2018. Link: https://doi.org/10.1371/journal.pone.0203629
2. Ding, L., Terwilliger, J., Parab, A., Wang, M., Fridman, L., Mehler, B., & Reimer, B. (2023). CLERA: A unified model for joint cognitive load and eye region analysis in the wild. ACM Transactions on Computer-Human Interaction. https://doi.org/10.1145/3603622
3. Egeonu, D., & Jia, B. (2025). A systematic literature review of computer vision-based biomechanical models for physical workload estimation. Ergonomics, 68(2), 139–162. https://doi.org/10.1080/00140139.2024.2308705

4. Papoutsakis, K., Bakalos, N., Fragkoulis, K., Zacharia, A., Kapetadimitri, G., & Pateraki, M. (2024). A vision-based framework for human behavior understanding in industrial assembly lines. arXiv preprint arXiv:2409.17356. https://arxiv.org/abs/2409.17356
5. Martinez-Cedillo, A. P., Gavrila, N., Mishra, A., Geangu, E., & Foulsham, T. (2025). Cognitive load affects gaze dynamics during real-world tasks. Experimental Brain Research, 243(4), 82. https://doi.org/10.1007/s00221-025-07037-4
6. Sevcenko, N., Appel, T., Ninaus, M. et al. Theory-based approach for assessing cognitive load during time-critical resource-managing human–computer interactions: an eye-tracking study. J Multimodal User Interfaces 17, 1–19 (2023). https://doi.org/10.1007/s12193-022-00398-y
7. Ekin, M., Krejtz, K., Duarte, C. et al. Prediction of intrinsic and extraneous cognitive load with oculometric and biometric indicators. Sci Rep 15, 5213 (2025). https://doi.org/10.1038/s41598-025-89336-y
8. Skaramagkas, V., Giannakakis, G., Ktistakis, E., Manousos, D., Karatzanis, I., Tachos, N. S., Tripoliti, E., Marias, K., Fotiadis, D. I., & Tsiknakis, M. (2023). Review of eye tracking metrics involved in emotional and cognitive processes. *IEEE Reviews in Biomedical Engineering, 16*, 260–277. https://doi.org/10.1109/RBME.2021.3066072
9. Nasri, M. (2025). Towards Intelligent VR Training: A Physiological Adaptation Framework for Cognitive Load and Stress Detection. arXiv preprint arXiv:2504.06461. https://doi.org/10.48550/arXiv.2504.06461
10. Cheng, J. C. P., Wong, P. K.-Y., Luo, H., Wang, M., & Leung, P. H. (2022). Vision-based monitoring of site safety compliance based on worker re-identification and personal protective equipment classification. *Automation in Construction, 139*, 104312. https://doi.org/10.1016/j.autcon.2022.104312
11. Agostinelli, T., Generosi, A., Ceccacci, S. et al. Validation of computer vision-based ergonomic risk assessment tools for real manufacturing environments. Sci Rep 14, 27785 (2024). https://doi.org/10.1038/s41598-024-79373-4

Exploring Vision Language Models for Egocentric Action Localization

Valentin Knoben[1,2], Julia Kramme[1], Björn Hein[1], Christian Wurll[1]

[1]Institute for Robotics and Autonomous Systems, Karlsruhe University of Applied Sciences
[2]Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology
{valentin.knoben, julia.kramme, bjoern.hein, christian.wurll}@h-ka.de

Abstract. Context-aware systems can support humans at work by automatically performing quality control, providing assistance, or generating instructions and documentation for latter use. However, the adaptation of such intelligent systems to custom use cases demands training data, expertise, and effort. With the dissem-ination of Vision Language Models (VLMs), recognition capabilities are becoming more accessible. We explore the use of readily available VLMs for understanding egocentric video footage of common manual tasks in production environments. Results demonstrate the feasibility of using VLMs in such contexts.

Keywords: Vision language model, action recognition, temporal localization

1 Introduction

In production-related environments, recognizing a worker's actions enables online quality checks [1] and real-time assistance [2]. Apart from simply understanding a performed task, extracting temporal information further supports the documentation of a workflow or the creation of digital instructions, e.g., for augmented reality-based manuals [3]. Research has particularly explored exocentric scene understanding, for instance, to enhance human robot interaction [4]. With the ongoing advancement of head-worn devices and their application in work environments [5], egocentric input data is becoming a potential source of information, too. Yet, reliably integrating such contextual understanding into a system requires expertise and effort in training a model on custom data or hiring a dedicated system integrator which are key barriers faced by Small and Medium-sized Enterprises (SMEs) in particular [6]. With the dissemination of Vision Language Models (VLMs) such as ChatGPT, Gemini, or Claude which can handle multimodal input, the barrier to entry has been reduced significantly in that regard. Therefore, we investigate the feasibility of using an off-the-shelf VLM to recognize and temporally locate common, production-related actions from videos. We focus on short clips showing a single action which might be performed as part of a more comprehensive workflow.

2 Methods

We recorded egocentric videos at different resolutions using the Microsoft HoloLens 2. Some VLMs can handle video input directly. However, internally the video input is usually sampled at 1fps which is insufficient for understanding short action clips. Therefore, we streamed footage to a near-edge unit at 30fps and extracted frames at 3fps. Frames were numbered to provide temporal anchors [7] (Fig. 1) and sent with a prompt asking the

VLM to return the action type and key frame(s) as structured JSON. The candidate actions were: pick-and-place, use wrench/Allen/screwdriver, press button, rotate switch, open/close element, fasten/loosen by hand. For each action we recorded three scenarios, provided ground truths for the correct action type and eligible key frames, and ran five trials each, resulting in 15 data points per action type. After initially testing different VLMs, we proceeded with gpt-4.1.

Fig. 1. Exemplary frames extracted from an egocentric image stream, fed to VLM. Returned json output: *{"action type"*: "rotate switch", *"key frames"*: "5, 6, 7"*}*

3 Results and Discussion

Fig. 2 illustrates the number of correct action recognitions and temporal localizations per action type, revealing high correct recognition counts alongside much more variable frame-level localizations. We identified two common failure modes. In case of incorrectly recognized action types, actions were simply confused, e.g., rotating a switch with pressing a button, potentially due to bad lighting conditions and both elements being visible in the scene. Causes of incorrect temporal localization are the strict classification as incorrect even when off by only 1 frame in ambiguous scenarios and the confusion between object visibility and its actual engagement. Apart from that, a temperature of around 0.3 and top p of 1 produced concise, low-hallucination outputs. Action recognition degraded noticeably below a resolution of 1504×846, while key-frame localization remained robust at lower resolutions, consistent with recognition needing fine spatial detail such as object geometry whereas localization relies on coarser motion cues. Increasing resolution beyond 1500 pixels in width gave diminishing returns, likely because VLMs tile large images to fixed sizes, fragmenting objects and limiting cross-tile context [8]. Possible leverage points for improvement are to perform hand- or object-based cropping of frames or concatenate multiple frames into a single image to reduce the number of pixels or frames to evaluate. However, this must carefully be balanced with recognition quality which could be aided by dynamic instead of fixed sampling rates.

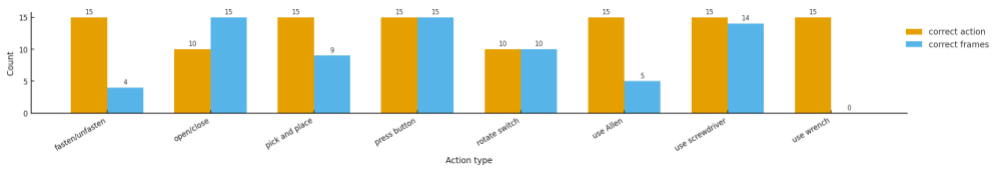

Fig. 2. Correct action recognitions (orange) and temporal localizations (blue) per action type

Acknowledgments. This research was conducted within The Cooperative Graduate School Accessibility through AI-based Assistive Technology (KATE) which is funded by the Ministry of Science, Research and the Arts of Baden-Württemberg (MWK) under Grant BW6 03.

References

1. Islam, M.R., Zamil, M.Z.H., Rayed, M.E., Kabir, M.M., Mridha, M., Nishimura, S., Shin, J.: Deep learning and computer vision techniques for enhanced quality control in manufacturing processes. IEEE Access (2024)
2. Li, W., Xu, A., Wei, M., Zuo, W., Li, R.: Deep learning-based augmented reality work in-struction assistance system for complex manual assembly. Journal of Manufacturing Systems **73** (April 2024) 307–319
3. Klimant, P., Kollatsch, C.: Concepts for creating augmented reality based technical docu-mentations for the maintenance of machine tools. International Journal on Interactive Design and Manufacturing (IJIDeM) **16**(2) (June 2022) 765–773
4. Ji, Y., Yang, Y., Shen, F., Shen, H.T., Li, X.: A survey of human action analysis in hri applications. IEEE Transactions on Circuits and Systems for Video Technology **30**(7) (2020) 2114–2128
5. Fang, W., Chen, L., Zhang, T., Chen, C., Teng, Z., Wang, L.: Head-mounted display aug-mented reality in manufacturing: A systematic review. Robotics and Computer-Integrated Manufacturing **83** (October 2023) 102567
6. Oldemeyer, L., Jede, A., Teuteberg, F.: Investigation of artificial intelligence in SMEs: a sys-tematic review of the state of the art and the main implementation challenges. Management Review Quarterly **75**(2) (June 2025) 1185–1227
7. Wake, N., Kanehira, A., Sasabuchi, K., Takamatsu, J., Ikeuchi, K.: Open-vocabulary action localization with iterative visual prompting (August 2024)
8. Dong, X., Zhang, P., Zang, Y., Cao, Y., Wang, B., Ouyang, L., Zhang, S., Duan, H., Zhang, W., Li, Y., Yan, H., Gao, Y., Chen, Z., Zhang, X., Li, W., Li, J., Wang, W., Chen, K., He, C., Zhang, X., Dai, J., Qiao, Y., Lin, D., Wang, J.: Internlm-xcomposer2-4khd: a pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Volume 37 of NIPS '24., Red Hook, NY, USA, Curran Associates Inc. (June 2025) 42566–42592

HelpMeWalk - A new digital process for orthoses production: Data processing, Workflow and Accuracy

Fabien Wilhelm[1], Dua Shahid[2], Maximilian Mock[3], Corentin Féry[4], Uwe Tronnier[5]

[1] Affiliation **fabien.wilhelm@hs-kl.de** [2] Affiliation **dua.shahid@hs-kl.de**
[3] Affiliation **maximilian.mock@hs-kl.de** [4] Affiliation **corentin.fery@fhnw.ch**
[5] Affiliation **uwe.tronnier@hs-kl.de**

Abstract. The *HelpMeWalk* project introduces a low-cost, magnetic field–based 3D acquisition system using sensor-embedded textile bandages for digital modeling of lower leg orthoses. The system captures approximately 300–500 spatial points within six seconds, providing data for a model-based surface reconstruction. A multi-step workflow – comprising error cleaning, alignment of measured points to a model surface and form adjustment – enables accurate reconstruction of individual limb geometries. Prototype evaluation demonstrated robust functionality and a final mean surface deviation of 2.89
mm. Remaining inaccuracies were mainly linked to sparse sensor coverage and sensor noise. Future development will focus on optimizing sensor placement protocols, mobile App integration, and machine learning enhanced error detection.

Keywords: Digital orthoses; magnetic tracking; 3D reconstruction; remote healthcare.

1 Introduction

The provision of orthopedic aids such as foot orthoses is still largely based on analog methods like plaster casting, which are labor-intensive, imprecise, and costly – especially when patient transport is needed. Inaccurate molds often require multiple adjustments, reducing comfort and patient adherence, and thereby compromising therapeutic success. With more than 300 custom orthoses produced annually per practice, there is strong demand for precise, reproducible, and efficient digital processes that improve both productivity and patient comfort.

The *HelpMeWalk* project aims to fill this gap by developing a cost-effective, user-friendly digital 3D measurement solution. Washable sensor-equipped bandages (“smart textiles”) capture a point cloud during manual corrective positioning, enabling model-based reconstruction of body structures with high accuracy from relatively few data points.

2 Method

3D point acquisition was achieved using a magnetic field tracking system, which determines the positions of three-axis magnetic field sensors. Trilateration of points is based on precomputed magnetic field maps and generates a point cloud of 300–500 points in under six

seconds. Different flexible PCB layouts were employed to effectively capture complex curvatures.

Surface reconstruction is model-based, mapping the acquired point cloud onto a normalized 3D surface model of an average lower leg. The process involves four main steps:

1. **Error Cleaning**: Measurement points are classified using geometric constraints derived from PCB design, comparing theoretical and measured distances. Points are scored by local relations and excluded if distances exceed physical limits [1].
2. **Rigid Pre-Registration and Size Adjustment**: The normalized model and measured point cloud are roughly aligned using artificial landmarks. Alignment includes translation and quaternion-based rotations. The normalized model is then scaled along the main anatomical axes of lower leg and foot to match patient-specific dimensions.
3. **Fine Registration**: An Iterative Closest Point (ICP) algorithm is used to refine alignment between the point cloud and the scaled surface model. This step benefits from error cleaning and a certain degree of similarity as prepared above [2], [3], [4].
4. **Shape Adjustment**: A Thin-Plate-Spline (TPS) algorithm [5] establishes nearest-neighbor correspondences from measured cloud points to the surface model and deforms the model accordingly. Nearest surface points are displaced toward their matched sensor point, while neighboring points are adjusted using a falloff function.

Results were post-processed using (1) a Windowed Sinc Smoothing to remove undetected outliers caused by random noise, (2) a voxel-based surface shrinking to correct for a 3 mm offset caused by the bandage thickness, and (3) a final smoothing to obtain a high-fidelity representation of the surface.

3 Results & Conclusion

Table 1. Geometric deviations between an optimal surface scan and process steps of a test reconstruction

Compare after	Size Adjustment	ICP Registration	TPS Adjustment	Post-process
Mean Distance	11,64 mm	4,14 mm	3,63 mm	2,89 mm
75th Percentile	17,04 mm	5,74 mm	5,29 mm	4,07 mm
Mean SE	196,31	26,21	22,04	14,11

A reconstruction test was performed which shows that the individual steps consistently decrease mean distance error down to 2,89 mm when comparing the results to a provided optical surface scan of the reference leg (Table 1). High deviations were mainly found in regions lacking sensors patches or when patches exhibit a consistent spatial offset due to systematic noise.

First results demonstrate the potential of the magnetic sensor patch approach of *HelpMeWalk* for accurate 3D reconstruction of lower limbs. Though broader validation across diverse patient groups and real-world conditions is still required. Accurate results depend on well-defined sensor placement protocols, as patch misplacement introduces early errors. Further research focuses on optimizing patch positioning, refining algorithms and incorporating variable foot angles between lower leg and foot as this is a treatment / costumer requirement. Improved data cleaning with Dynamic Graph CNNs and model selection based on individual patient characteristics are currently being investigated to improve overall accuracy.

References

1. Rakotosaona, M.-J.; La Barbera, V.; Guerrero, P.; Mitra, N. J.; Ovsjanikov, M. (2019): "PointCleanNet: Learning to Denoise and Remove Outliers from Dense Point Clouds." *Computer Graphics Forum*, 38(1), 275–286.
2. Besl, P. J.; McKay, N. D. (1992): "A Method for Registration of 3-D Shapes." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2), 239-256.
3. Zinßer, Timo; Schmidt, Jochen; Niemann, Heinrich (2005): "Point Set Registration with Integrated Scale Estimation." *Int. Conf. on Pattern Recognition and Image Processing (PRIP 2005)*, 116-119.
4. Gang, Wang; Qiangqiang, Zhou; Yufei, Chen (2017): "Robust Non-Rigid Point Set Registration Using Spatially Constrained Gaussian Fields." *IEEE Transactions on Image Processing*, 26(4), 1759-1769.
5. Chui, H., & Rangarajan, A. (2003). "A new point matching algorithm for non-rigid registration." *Computer Vision and Image Understanding*, *89*(2-3), 114-141.

Optimizing Decision Parameters of Humanoid Robots using Deep Reinforcement Learning

Richard Pufe
Department of Electrical Engineering, Medical Engineering and Computer Science, Offenburg University
rpufe@stud.hs-offenburg.de

Abstract. This work investigates the use of deep reinforcement learning to enable humanoid Nao robots in the RoboCup 3D Soccer Simulation to autonomously decide when to switch between complex behaviors. Two main experiments were conducted. In the first, an agent was trained to learn the optimal moment to transition from walking towards the ball to executing a kick. The robot was randomly initialized at varying distances and orientations relative to the ball and trained using Proximal Policy Optimization to maximize accuracy in kicking the ball towards a target after approaching it. The resulting models achieved strong performance on par with the handcrafted baseline in simulated matches. The second experiment extended this setup by allowing the agent to also determine a favorable pre-kick position round the ball before deciding to switch. Despite the richer decision space, the resulting models performed significantly worse than the baseline, indicating the increased difficulty of jointly learning spatial positioning and timing.

Keywords: Deep Reinforcement Learning; Behavior Switching; Humanoid Robots.

1 Introduction

In simulated humanoid soccer, coordinated decision-making between locomotion and ball interaction is essential for effective play. Conventionally, the timing for transitioning from approaching to kicking the ball is determined through manually tuned thresholds or heuristic rules, which restrict adaptability to diverse situations. This work explores an end-to-end learning approach in which a robot autonomously learns the appropriate timing and later also the positioning, necessary to execute a successful kick.

Experiments were conducted in the RoboCup 3D Soccer Simulation SimSpark [1] using the Nao humanoid robot model [2]. The training task required the robot to approach a stationary ball and kick it towards a fixed target position. The robot was randomly initialized at various distances and orientations relative to the ball to encourage generalization. To stabilize training, the robot's spawning area was gradually increased. Initially, the robot was initialized within a narrow angular sector behind the ball to simplify early exploration. As training progressed, the initialization range expanded to the full 360°, requiring the policy to handle complex approach directions and orientation corrections.

Training utilized the Proximal Policy Optimization (PPO) [3] algorithm with continuous action spaces. Each control cycle, the agent received a 170-dimensional observation vector containing information from onboard sensors and internal states. This included joint angles and speeds, accelerometer and gyroscope readings, torso pose, locomotion steering outputs, foot force sensors and estimates of the robot's linear and angular velocities. Additionally, ball-related features such as its relative position and velocity were provided to inform spatial awareness. All features were normalized to a fixed range to stabilize learning and ensure comparable magnitudes across dimensions. The reward function was designed to encourage the robot to approach the ball efficiently and to switch to the kicking behavior at the appropriate distance to maximize the accuracy of the kick towards the target position.

Two learning configurations were investigated: In this first configuration, the model's output represented a binary decision indicating whether to switch from walking to kicking. As long as the network output remained below the switching threshold, the robot continued to execute a fixed walking behavior towards the ball. Once the model signaled a switch, the robot transitioned to a predefined kicking behavior. The learning process thus focused solely on identifying the optimal timing for the transition, without any influence on the locomotion or kick mechanics themselves. This setup allowed the policy to concentrate on temporal decision-making based on the evolving spatial relationship between the robot and the ball. The second configuration extended the output space of the model to include a relative position and approach angle around the ball in addition to the binary switching signal. These continuous outputs were interpreted as a desired offset position towards which the robot navigated using its walking

behavior. However, the model was free to trigger the switch to the kicking behavior at any time and it was not required to reach the proposed position before switching. Thus, as long as the switching output remained inactive, the robot continued to walk towards the indicated position and as soon as the model decided to switch, the transition to the kicking behavior occurred immediately. This setup allowed the policy to dynamically adjust both positioning intent and switching timing within the same episode, coupling spatial and temporal reasoning in a flexible but more complex learning process that proved difficult to optimize effectively.

2 Conclusion

The experiments demonstrated that deep reinforcement learning can learn to imitate handcrafted decision mechanisms for switching between walking and kicking behaviors in simulated humanoid soccer. In the first experiment, where the agent learned only the switching decision, the models achieved strong performance during training but still slightly lagged behind the handcrafted baseline in terms of overall accuracy. However, the implementation of a curriculum learning strategy, gradually increasing the range of initialization angles around the ball, proved essential to achieving near-baseline performance during full simulated matches (Table 1). This curriculum allowed the agent to generalize across a wider range of approach directions, improving robustness. One notable observation was that the learned model typically decided to switch earlier, requiring fewer control cycles before executing a kick compared to the baseline, which partially compensated for its slightly lower target accuracy.

Table 1. Simulated match outcomes. Each learned policy (Team A) competed against the handcrafted baseline (Team B) over multiple games. The baseline entry represents self-play, where both teams used the handcrafted decision mechanism. *Goals A/B* are per-match means and *Wins A/B* and *Ties* are counts over Games.

Version	Games	Goals A	Goals B	Wins A	Wins B	Ties
Baseline	400	2.788	2.858	158	159	83
Switch Model	500	2.938	3.022	182	208	110
Positioning Model	500	1.684	4.424	13	456	31

The second experiment, which extended the learning task to include both positioning and switching, introduced significantly higher complexity. None of the trained models were able to reliably learn both aspects well enough, resulting in performance that fell well below the baseline in both training evaluations and match simulations (Table 1). This highlights the increased difficulty of jointly learning spatial positioning and temporal decision-making within a single policy. Future improvements may require refined or extended curriculum strategies and more targeted fine-tuning to help the agent learn effective positional reasoning alongside switching behavior.

References

[1] J. Boedecker and M. Asada, "SimSpark - concepts and application in the RoboCup 3 d soccer simulation league," 2008. Accessed: Jul. 31, 2025. [Online]. Available: https://www.semanticscholar.org/paper/SimSpark-%E2%80%9 3-Concepts-and-Application-in-the-RoboCup-Boedecker-Asada/872b60cbbf89bda70488d3089958140 740a60a14.

[2] R. Gelin, "NAO," in *Humanoid Robotics: A Reference*, Springer, Dordrecht, 2019, pp. 147–168, ISBN: 978-94-007-6046-2. DOI: 10.1007/978-94-007-6046-2_14. Accessed: Jul. 31, 2025. [Online]. Available: https://link.springer. com/rwe/10.1007/978-94-007-6046-2_14.

[3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, *Proximal policy optimization algorithms*, Aug. 28, 2017. DOI: 10.48550/arXiv.1707.06347. arXiv: 1707.06347[cs]. Accessed: Mar. 8, 2025. [Online]. Available: http://arxiv.org/abs/1707.06347.

Towards DQN Reinforcement Learning for energy management for bidirectional charging of EV's

Rainer Gasper[1], Michael Quarti[1], Nick Abermeth[1], Yannik Heizmann[1], Joshua Ruf[1], Markus Portugal[1], Bennet Märtin[1]

[1] Hochschule Offenburg
rainer.gasper@hs-offenburg.de

Abstract. This study applies Deep Q-Network (DQN) reinforcement learning to optimize bidirectional EV charging in a microgrid with dynamic pricing and renewable energy. The environment includes an EV, wind turbine, stationary battery, flexible household loads, and grid connection. DQN agents learn to minimize energy costs by charging during low-price periods and discharging during high-price windows. Simulations across four scenarios show improved cumulative rewards and grid efficiency. Future work will address stochastic elements, realistic EV availability, and continuous action spaces to enhance adaptability and performance in real-world applications.

Keywords: Reinforcement Learning; DQN; MicroGrid, Bidirectional Charging.

1 Introduction

The integration of electric vehicles (EVs) into microgrids presents a transformative opportunity for local energy systems. The bidirectional charging of EVs enables vehicle-to-grid (V2G) services, which leads to enhancing grid stability, reducing peak loads, and supporting renewable energy integration. However, managing bidirectional charging in dynamic environments like microgrids is complex due to fluctuating energy demand, supply, and user behavior.

Deep Q-Network (DQN) reinforcement learning offers a promising strategy to address this challenge. By learning optimal policies through interaction with the environment, DQN can manage charging and discharging decisions, adapting to real-time conditions and maximizing long-term rewards such as cost savings.

2 DQN Reinforcement Learning

DQN combines Q-learning with deep neural networks to approximate the optimal policy function. The DQN is implemented with experience replay, where the transitions (state, action, reward, next state) are stored in a memory buffer and sampled randomly during training [2]. Also, a target network is used to compute target values [3]. Both technics improve the learning process by reducing oscillations and breaking correlation between sequential data [2,3]. For the exploitation and exploring trade-off, we implemented Ɛ-greedy-strategy with decreasing Ɛ. The Q-learning allows only discrete actions (e.g. charge, discharge, buy or sell) based on the current state of the environment. The learning takes place in the agent, he receives rewards based on its actions and updates its policy to maximize cumulative future rewards, the cost savings.

The environment consists of the microgrid and the reward function. The microgrid itself consists of an EV with bidirectional charging capability, a wind turbine as a renewable energy source, a stationary battery, and households as consumers with flexibilities to shift their power consumption [3]. Also, the grid connection point with the dynamic electricity pricing is included, which acts as a source (buying) or sink (selling). The data for the microgrid are taken from [3]. We neglected the uncertainties in the microgrid, therefore the environment is fully deterministic. The reward function calculates the energy costs (multiplied by minus one).

Fig. 1. Test results for net profit over 100 days for baseline batterie, baseline grid, price action and shift action

3 Simulation Results

In the simulation, we compare four different scenarios to evaluate the results. The first scenario is the baseline, where no batteries are used and every excess power is sold, and vice versa. The second scenario includes the stationary battery, where the excess power is stored in the batterie, and a power gap leads to discharging of the battery. Only if the limits of the battery are reached, the grid power is used. In the third scenario the agent minimized the price, and in the fourth scenario, the agent can also shift the power demand of some consumers. The initial results, see Fig. 1, for 150 households show that the agents learn to discharge during high-price periods and charge during low-price windows. Compared to baseline scenarios, the DQN-agents achieve higher cumulative rewards.

4 Conclusion and Outlook

The first results show already consistent behavior of the DQN – agents. Future work will also include stochastics in the renewable energy, the price forecast, and in the consumer behavior. This includes the availability of the EV battery. Also, the influence of different reward functions has to be investigated further. Another point for future work is to implement agents with continuous actions for power distribution.

References

1. T. A. Nakabi, P. Toivanen, Deep reinforcement learning for energy management in a microgrid with flexible demand, Sust. Energy, Grids and Networks, Volume 25, 2021, https://doi.org/10.1016/j.segan.2020.100413

2. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., et al. (2013). Playing Atari with Deep Reinforcement Learning. CoRR, abs/1312.5602.
3. Mnih, V., Kavukcuoglu, K., Silver, D. et al., Human-level control through deep reinforcement learning. Nature 518, 529–533 (2015). https://doi.org/10.1038/nature14236

Usability Study of an Intelligent Robotic Walker for Lower-Limb Motor Training

Andreas Zachariae[1,2], Christian Wurll[2], and Björn Hein[2,1]

[1] Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology
[2] Institute for Robotics and Autonomous Systems, Karlsruhe University of Applied Sciences
{andreas.zachariae, christian.wurll, bjoern.hein}@h-ka.de http://www.h-ka.de/iras

Abstract. The RoboTrainer is an intelligent robotic walker that provides force-controlled lower-limb motor training. This short paper presents the results of a usability study using the standardized System Usability Scale (SUS) question-naire. The average SUS score from 28 healthy participants was 76.6 ± 12.5. We compared this score with a meta-analysis of 67 robotic rehabilitation devices, demonstrating that the RoboTrainer's usability exceeds the field's average. These findings suggest the technology has strong potential for successful adoption.

Keywords: System Usability Scale, Human-Robot Interaction, Robotic Walker

1 Introduction

Lower-limb motor performance declines with increasing age due to muscle loss (35% be-tween the ages of 40 and 80 [1]). This can lead to an increased risk of injury from falls [2]. Additionally, over 20% of nursing home residents have mild cognitive impairment [3], and 60% of these individuals will develop dementia during their lifetime [4]. Studies show that interventions involving motor and cognitive stimulation can maintain or even improve these functions [5][6]. In this context, individualized and adaptive training programs are essential, as they can provide appropriate challenges to ensure effective training. Given the aging society [7] and the shortage of skilled healthcare workers [8], automating the training process with easy-to-use devices is essential to address these issues.

The RoboTrainer, an intelligent robotic walker currently in development, addresses this problem by providing force-controlled lower-limb motor training with additional cognitive challenges [9]. The RoboTrainer uses machine learning to automatically assess performance and adapt task difficulty to provide individualized training [10].

High usability is important if a device is to be adopted in care homes and used for daily training. It must be intuitive to use for both patients and clinical staff. Poor usability can lead to user frustration, high error rates, patient rejection, and ultimately, abandonment of the technology, regardless of its therapeutic potential.

The contributions of this paper are the results of a usability study:

1. Presentation of the System Usability Scale (SUS) evaluation of the RoboTrainer.
2. Comparison of the results with other robotic rehabilitation devices.

2 Method

The usability study was part of a larger technical feasibility study aimed at collecting data and testing the automatic assessment algorithms. A total of 28 participants (9 female, 19 male; mean age 34.7 ± 12.8 years) were recruited for this study. The inclusion criterion

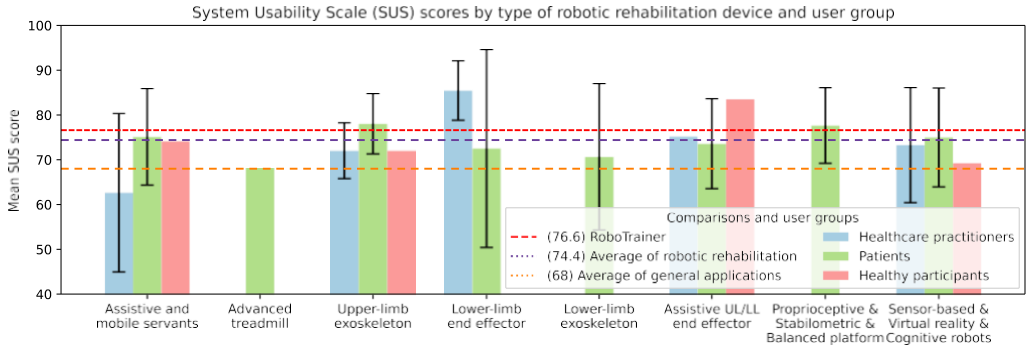

Fig. 1. The SUS score of RoboTrainer compared with other robotic rehabilitation devices [14].

was the absence of cognitive or mobility impairments. Each participant completed a series of 20 different tasks with RoboTrainer for one hour. After the session, usability was quantified using the standardized 10-item SUS questionnaire [11]. The SUS is a validated questionnaire that provides a global score of perceived usability from 0 to 100.

3 Results

The reported SUS score for RoboTrainer is 76.6 ± 12.5, which is considered "good" usability and belongs to the top 22nd percentile [12]. This indicates a high level of user acceptance, ease of use, and system consistency. The average score for general industry applications is 68 ± 12.5 [13].

To contextualize these findings, we compared our results to a recent meta-analysis by Nicora et al. (2025), which analyzed 67 studies on robotic rehabilitation devices [14]. The average SUS score across all devices and user groups was 74.4 ± 11.3 as shown in Figure
1. The RoboTrainer belongs to the device category of lower-limb end effectors and was evaluated with healthy participants. This device category includes a total of six studies with a mean score of 85.4 ± 6.6. The most comparable device is the FriWalk robotic walker, which recieved a score of 52.8 by 42 elderly patients and 68.0 by one clinical professional [15]. Another is the SWalker, which achieved 91.3 by two physiotherapists [16]. Furthermore, our healthy user score of 76.6 can be compared to the meta-analysis average for all studies involving healthy participants (n = 4), which was 74.0.

The primary limitation of this technical pilot study is its use of a healthy user group that does not represent the final target population of older adults with potential impair-ments. This limitation is also reflected in the qualitative feedback, as some participants found the question about "frequent use" (item 1) difficult to answer. However, the over-whelmingly positive score confirms the system's core intuitiveness and ease of interaction.

4 Conclusion

This paper presents a usability evaluation of the RoboTrainer as a first step towards its clinical application. The resulting SUS score of 76.6 ± 12.5 from healthy partici-pants exceeds the average usability scores for both general industry applications and other rehabilitation devices. Future work will focus on validating these promising results with the target population of older adults and on further developing the RoboTrainer's individualization algorithms.

Acknowledgement. This work is supported by the Cooperative Graduate School Accessibility through AI-based Assistive Technology (KATE), funded by the Ministry of Science, Research and Arts of Baden-Wuerttemberg (MWK) under Grant BW6_03.

The user study conducted with RoboTrainer was approved by the ethics committee of the Karlsruhe Institute of Technology under registration number A2025-083.

References

1. Akima, H., Kano, Y., Enomoto, Y., Ishizu, M., et al.: Muscle function in 164 men and women aged 20-84 yr:. Medicine and Science in Sports and Exercise (February 2001) 220–226
2. Pfortmueller, C.A., Lindner, G., Exadaktylos, A.K.: Reducing fall risk in the elderly: risk factors and fall prevention, a systematic review. Minerva Medica **105**(4) (August 2014) 275–281
3. Chen, P., Cai, H., Bai, W., Su, Z., et al.: Global prevalence of mild cognitive impairment among older adults living in nursing homes: a meta-analysis and systematic review of epi-demiological surveys. Translational Psychiatry **13**(1) (March 2023) 88
4. Busse, A., Angermeyer, M.C., Riedel-Heller, S.G.: Progression of mild cognitive impair-ment to dementia: a challenge to current thinking. British Journal of Psychiatry **189**(5) (November 2006) 399–404
5. Heyn, P., Abreu, B.C., Ottenbacher, K.J.: The effects of exercise training on elderly persons with cognitive impairment and dementia: A meta-analysis. Archives of Physical Medicine and Rehabilitation **85**(10) (October 2004) 1694–1704
6. Ngandu, T., Lehtisalo, J., Solomon, A., et al.: A 2 year multidomain intervention of diet, exercise, cognitive training, and vascular risk monitoring versus control to prevent cognitive decline in at-risk elderly people (FINGER): a randomised controlled trial. The Lancet **385**(9984) (June 2015) 2255–2263
7. United Nations: Department Of Economic And Social Affairs: Population Division: World Population Ageing 2019. UNITED NATIONS, [S.l.] (2021) OCLC: 1289919896.
8. Statistisches Bundesamt: Pflegekräftevorausberechnung
9. Stogl, D., Armbruster, O., Mende, M., Hein, B., Wang, X., Meyer, P.: Robot-Based Training for People With Mild Cognitive Impairment. IEEE Robotics Autom. Lett. **4**(2) (2019) 1916–1923
10. Zachariae, A., Štogl, D., Hein, B., Wurll, C., Krell-Roesch, J., Woll, A.: Towards Automated, Individualized, and Adaptive Lower-Limb Motor Training with a Force-Controlled Robotic Walker. In: 2025 IEEE International Conference on Advanced Robotics and its Social Impacts (ARSO), Osaka, Japan, IEEE (July 2025) 146–151
11. Brooke, J.: SUS-A quick and dirty usability scale. In: Usability evaluation in industry. Volume 194. Taylor and Francis, London (1996) 189–194
12. James R. Lewis, Jeff Sauro: Item Benchmarks for the System Usability Scale - JUX. The Journal of User Experience **13**(3) (May 2018) 158–167
13. Sauro, J.: Quantifying the user experience: practical statistics for user research. Morgan Kaufmann, Place of publication not identified (2016)
14. Nicora, G., Santangelo, G., Marzero, G., Dagliati, A., Aprile, I., Quaglini, S.: Exploring Usability Standards in Robotic Rehabilitation: A Meta-Analysis of System Usability Scale Scores. In: 2025 International Conference On Rehabilitation Robotics (ICORR), Chicago, IL, USA, IEEE (May 2025) 476–482
15. Pérez-Rodríguez, R., Moreno-Sánchez, P.A., Valdés-Aragonés, M., Oviedo-Briones, M., et al.: FriWalk robotic walker: usability, acceptance and UX evaluation after a pilot study in a real environment. Disability and Rehabilitation: Assistive Technology **15**(6) (August 2020) 718–727
16. Costa, V., Ramirez, O., Perea, L., Velasquez, A., et al.: Development and Clinical Validation of a Rehabilitation Platform for Hip Fracture in Elderly Population. IEEE Transactions on Neural Systems and Rehabilitation Engineering **30** (2022) 1340–1349