Editorial: The Second Workshop on Explainable Artificial Intelligence for the medical domain - EXPLIMED@ECAI2025

Gabriella Casalino[1], Giovanna Castellano[1], Katarzyna Kaczmarek-Majer[2,3], Raffaele Scaringi[1,*] and Gianluca Zaza[1,*]

[1]*Computer Science Department, University of Bari Aldo Moro Bari, Italy*

[2]*Systems Research Institute Polish Academy of Sciences – Warsaw, Poland*

[3]*University of Ostrava, Institute for Research and Applications of Fuzzy Modeling, NSC IT4Innovations – Ostrava, Czech Republic*

Abstract

The 2025 Second Workshop on Explainable Artificial Intelligence for the Medical Domain (EXPLIMED) was held in conjunction with the 28th European Conference on Artificial Intelligence (ECAI 2025) in Bologna, following the success of its inaugural edition at ECAI 2024 in Santiago de Compostela. This year, the workshop reaffirmed its impact with the acceptance of 15 high-quality papers, carefully selected through a rigorous review process. EXPLIMED served as a forum for leading researchers and practitioners in Artificial Intelligence to explore the most recent advancements and best practices in Explainable AI (XAI) for healthcare. The event fostered discussions on current trends, ongoing research initiatives, and novel approaches to explainability in medical applications. By emphasizing a comprehensive and multidisciplinary perspective, the workshop highlighted how explainable methodologies can contribute to improving clinical decision-making and patient care.

Keywords

Explainable Artificial Intelligence, Trustworthy AI, Clinical Decision Support Systems, Healthcare, Bioinformatics

1. Introduction

The EXPLIMED workshop serves as a key venue for advancing Explainable Artificial Intelligence (XAI) in the medical domain, highlighting emerging research directions, methodological innovations, and applied studies. As AI technologies become increasingly embedded in healthcare decision-making, the workshop provides a collaborative space where scholars, clinicians, and policymakers can share perspectives on strengthening transparency, interpretability, and trust in medical AI systems.

EXPLIMED 2025 - Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy

*Corresponding author.

gabriella.casalino@uniba.it (G. Casalino); giovanna.castellano@uniba.it (G. Castellano); K.Kaczmarek@ibspan.waw.pl (K. Kaczmarek-Majer); raffaele.scaringi@uniba.it (R. Scaringi); gianluca.zaza@uniba.it (G. Zaza)

000-0003-0713-2260 (G. Casalino); 0000-0002-6489-8628 (G. Castellano); 0000-0003-0422-9366 (K. Kaczmarek-Majer); 0000-0001-7512-7661 (R. Scaringi); 0000-0003-3272-9739 (G. Zaza)

Explainable Artificial Intelligence (XAI) has become a cornerstone of research in the medical domain, where algorithmic decisions often have life-critical consequences. In this context, predictive accuracy alone is insufficient: clinicians and patients alike must be able to understand, validate, and trust the reasoning behind AI systems before their recommendations can be integrated into care pathways [1, 2]. This need is particularly pressing in light of recent European regulations, such as the AI Act, which mandate transparency and accountability in automated data analysis processes [3].

Among the different areas of application, medical imaging has become one of the most widespread domains where explainability is indispensable. Deep learning systems are increasingly employed in radiology, pathology, and related disciplines, yet their black-box nature raises concerns about clinical reliability. Physicians must understand not only what a model predicts but also why. Techniques such as saliency maps, counterfactual explanations, and prototype-based reasoning have been developed to bridge the gap between raw predictions and clinically meaningful visual evidence [4, 5]. However, imaging is not the only relevant data type: medical information often also comes in the form of tabular data or time series, such as those collected from monitoring devices and sensors, which demand tailored explainability methods [6].

These efforts connect to a broader movement toward rigorous validation of explanations. Early XAI demonstrations were often illustrative rather than systematic, but the field is now converging on measurable criteria such as fidelity, plausibility, stability, and clinical relevance. Establishing robust metrics and evaluation frameworks is vital to ensure that explanations do more than appear intuitive—they must demonstrably support medical decision-making and be reliable across contexts [7, 8].

At the same time, researchers are increasingly advocating for human-centered XAI, which places physicians and patients at the core of the design and evaluation process. Explanations must align with clinical workflows, cognitive expectations, and the practical constraints of medical environments, rather than being designed solely from a technical standpoint [9, 10, 11].

The recent rise of generative AI (GenAI) introduces both opportunities and challenges. Large language models are powerful tools for generating natural language explanations, potentially making technical outputs more accessible to clinicians. Yet their opaque mechanisms and risk of hallucination represent the antithesis of explainability. A promising path forward is to use GenAI as a narrative layer built upon inherently interpretable models, combining the communicative strength of LLMs with the reliability of transparent reasoning [12, 13].

In parallel, neuro-symbolic AI is gaining traction as another hybrid paradigm. By combining data-driven learning with symbolic reasoning, these approaches aim to retain the predictive power of neural models while embedding rule-based structures that align with human reasoning. This integration holds the potential to make complex AI systems both more interpretable and more trustworthy in clinical practice [14].

All these research directions converge on the broader vision of Trustworthy AI, which encompasses not only explainability but also fairness, accountability, and transparency [15, 1]. In this context, the EXPLIMED workshop provides a forum that brings together researchers working on XAI for the medical domain from multiple perspectives, some of which are highlighted in this editorial. Beyond presenting scientific advances, the workshop also serves as a meeting point for fostering collaboration, networking, and community building around the shared goal

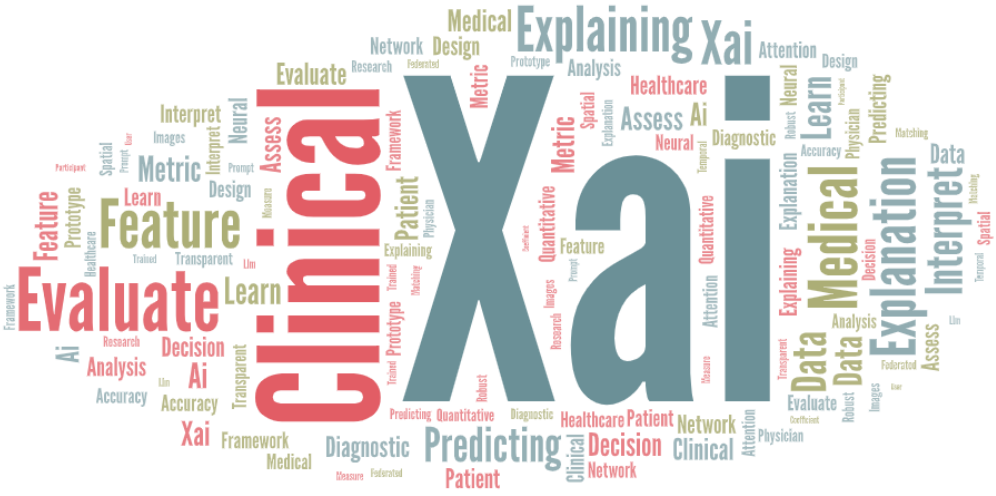

Figure 1: Wordcloud showcasing the most frequently occurring terms in the abstract of the workshop documents.

of making medical AI more transparent, reliable, and human-centered.

2. Workshop's Contibutions

We received 24 submissions for the EXPLIMED workshop, and after a thorough review process, 15 were accepted for presentation. Researchers from 17 countries (India, Belgium, Poland, Netherlands, Switzerland, Spain, Russia, United Kingdom, United States, Pakistan, Greece, Tunisia, Canada, Ireland, Italy, France, Algeria) submitted a paper to the workshop. This highlights the international significance of the topic Explainable Artificial Intelligence and underscores ECAi conference as a pivotal event for discussing AI-related issues, not only for European researchers but for the global academic community.

As shown in Figure 1, the workshop contributions cover a wide spectrum of topics, each addressing essential dimensions of XAI in the medical domain. Terms such as evaluate, quantitative, metric, and assess emphasize the importance of developing rigorous methods to evaluate explanations, a key requirement for building trust in AI systems. The recurrent use of patient and physician highlights the central role of end-users in shaping methodologies that prioritize human-centered analysis and explanation. The prominence of accuracy underlines that explainability must go hand in hand with models that are robust and effective in practice. Other relevant terms, including prototype, neural network, feature, spatial, and federated, reflect the variety of explainable algorithmic approaches currently explored in medical AI. The frequent mention of data and images confirms that at the foundation of both AI and explainability lie high-quality datasets, without which no meaningful progress can be achieved. Finally, the presence of framework signals the urgent need for community-accepted standards that can be shared and reused, ensuring consistency and comparability in explainability research. Not

surprisingly, the most prominent terms, clinical and XAI, underscore the workshop's dual focus on advancing explainability while addressing the specific challenges of clinical practice.

2.1. Session 1

EXPLIMED opened with an inspiring keynote delivered by Prof. *Alberto Fernández Hilario*, Full Professor of Computer Science and Artificial Intelligence at the University of Granada, titled *Trust in Practice: Rethinking Explainability and Fairness for Responsible AI*. The keynote addressed the growing deployment of Artificial Intelligence in high-stakes domains where reliability, transparency, and fairness are indispensable. Prof. Fernández Hilario presented recent advances from his research on counterfactual explanations and fairness-aware preprocessing techniques—two approaches that extend beyond predictive accuracy to ensure AI systems are both interpretable and equitable. Counterfactual explanations provide actionable insights by demonstrating how changes in input can alter model predictions, while fairness-oriented preprocessing tackles bias directly at the data level, underscoring the critical role of data-centric AI in building trustworthy systems. Drawing on clinical decision-making as a representative high-risk scenario, the keynote illustrated how explanations must remain understandable and actionable, and how fairness in patient data is central to responsible medical AI. Through concrete case studies, Prof. Fernández Hilario emphasized the necessity of designing AI systems guided by the principles of Trustworthy AI, and highlighted the value of human-in-the-loop strategies for aligning automation with expert oversight.

The talk demonstrated how methodological innovations can be effectively translated into practice, bridging the gap between technical progress and responsible deployment, especially in sensitive domains such as healthcare.

2.2. Session 2

The paper *Federated Inductive Logic Programming for Explainable Artificial Intelligence* [16] presents a novel framework, called FILP, that integrates Inductive Logic Programming (ILP) into a Federated Learning (FL) setting to address two persistent challenges in healthcare AI: data privacy and model explainability. Unlike conventional FL approaches that rely on black-box neural models, FILP leverages the symbolic and declarative nature of ILP to generate interpretable rules from distributed clinical datasets without exposing raw patient data. Each client independently learns logical theories guided by background knowledge, and a consensus mechanism aggregates them into a global interpretable model. The framework is instantiated with two ILP engines, Andante (Progol-based) and Popper (ASP-based), demonstrating its generality across symbolic paradigms. FILP is applied to a real-world clinical scenario—predicting post-intubation complications—and also evaluated on established ILP benchmarks. The proposed methodology emphasizes privacy-preserving, human-understandable reasoning, bridging the gap between transparency and effectiveness in medical AI.

The paper *Massive Activations in Graph Neural Networks: Decoding Attention for Domain-Dependent Interpretability* [17] investigates an underexplored phenomenon in edge-featured Graph Neural Networks (GNNs): the occurrence of Massive Activations (MAs) in attention layers. The authors develop methods to detect and characterize these extreme activations,

showing that they are not random anomalies but encode domain-relevant signals. Through empirical analysis on molecular and protein datasets, they demonstrate how MAs can act as natural attribution indicators, shedding light on the interplay between attention mechanisms and graph structures. The study provides a framework for post-hoc interpretability, offering new insights into the internal dynamics of attention-based GNNs and their connection to domain knowledge

The paper *Comparative Plausibility Evaluation of Heatmaps for Vision Transformers in Digital Mammography* [18] addresses the challenge of evaluating explainability methods for Vision Transformers (ViTs) applied to medical imaging. While heatmaps are widely used to visualize which image regions contribute most to model predictions, their reliability and clinical plausibility remain uncertain. The authors introduce a framework that uses the concept of "plausibility," defined as the alignment of generated explanations with ground truth, to objectively assess heatmap quality. The study explores different explainability techniques, including CAM- and attention-based methods, and compares their ability to highlight diagnostically relevant areas in mammography images. By formalizing plausibility metrics and applying them to synthetic mammogram data, the paper contributes to establishing a systematic approach for evaluating XAI methods in medical imaging, supporting their use in clinical and regulatory contexts.

The paper *High Cost, Low Trust? MSA-PNet Fixes Both for Medical Imaging* [19] introduces a novel deep learning framework designed to improve efficiency and explainability in ultrasound-based disease diagnosis. Traditional architectures such as EfficientNetB7 and VGG19 suffer from high computational demands and limited transparency, which hinder clinical adoption. To address these challenges, the authors propose MSA-PNet, a Multi-Scale Attention-Enhanced Prototype Network that combines adaptive feature fusion, spatial attention mechanisms, and a region of interest (ROI) segmentation branch to localize tumor regions with precision. A key element of the architecture is its prototype-based explainability module, which grounds predictions in class-specific prototypical patterns and supports case-based reasoning. The approach is motivated by the need for models that are not only accurate but also interpretable and efficient, enabling real-time application in clinical workflows.

The paper *CRISP-NAM: Competing Risks Interpretable Survival Prediction with Neural Additive Models* [20] addresses the challenge of modeling survival data in healthcare where patients may face multiple competing risks. Traditional statistical methods like Cox models provide interpretability but struggle with nonlinearities, while deep learning approaches improve accuracy but lack transparency. To bridge this gap, the authors propose CRISP-NAM, an extension of Neural Additive Models designed to jointly estimate cause-specific hazards while maintaining feature-level interpretability. The framework allows each feature to contribute independently to different risks, enabling visualization of non-linear effects and generating shape functions to understand covariate influences. This inherently interpretable approach supports compliance with regulatory requirements and enhances trust in predictive models for clinical decision-making

2.3. Session 3

The paper *An Interpretable Prototype Parts-based Neural Network for Medical Tabular Data* [21] introduces a novel neural architecture, MEDIC, specifically designed for clinical tabular

datasets. Inspired by prototype-based networks in computer vision, the model adapts the idea of comparing new cases to representative prototypes but redefines "parts" for structured data rather than spatial features. MEDIC combines discretization of clinical variables, learnable masks for grouping features, and prototypical representations aligned with real patient records. This approach allows predictions to be expressed as comparisons with human-understandable clinical cases, facilitating transparency and interpretability. The framework aims to bridge predictive performance and explainability by grounding decisions in medical concepts, making it suitable for supporting trust in healthcare decision-making.

The paper *Using Design Thinking for Explainable AI: A Case Study Predicting the Start of the Palliative Phase in Patients with COPD or Heart Failure* [22] explores how the Design Thinking methodology can be applied to create user-centered explainable AI (XAI) solutions in healthcare. The study focuses on supporting healthcare professionals in recognizing the palliative phase in patients with chronic obstructive pulmonary disease (COPD) or heart failure. Rather than prioritizing technical explanations, the work emphasizes including healthcare practitioners in the design cycle to ensure the XAI representation aligns with their workflows and decision-making needs. The case study highlights the ideation, prototyping, and testing phases, aiming to generate representations that help professionals understand AI outputs while maintaining responsibility for clinical decisions. This contributes to shaping a structured, iterative design process for XAI systems tailored to medical contexts.

The paper *Evaluation of Explainable AI by Medical Experts: a Survey of the Existing Approaches* [23] presents a structured review of how Explainable Artificial Intelligence (XAI) techniques are evaluated in the medical domain by practitioners. It highlights the lack of standardized frameworks for assessing XAI usability and trustworthiness in clinical contexts. The authors analyze existing evaluation methodologies, pointing out recurring issues such as insufficient reporting of study details, limited use of statistical validation, and a tendency to test XAI tools in isolation. The survey also identifies gaps in the range of properties assessed, with an overemphasis on usefulness and clinical relevance, while neglecting broader aspects of explanation quality. To address these shortcomings, the paper proposes a set of guidelines and recommendations to improve the design, execution, and reporting of XAI evaluation studies, aiming to ensure more rigorous and meaningful integration of XAI into healthcare practice

The paper *MRxaI: Black-Box Explainability for Image Classifiers in a Medical Setting* [24] addresses the challenge of evaluating explainability methods in medical imaging. Traditional metrics such as the Sørensen–Dice coefficient, Jaccard Index, and Hausdorff Distance only partially capture clinically relevant aspects like size, location, and continuity of explanations. To overcome these limitations, the authors introduce the Penalized Dice Coefficient (PDC), a novel metric that integrates spatial alignment, area similarity, and fragmentation into a single measure tailored to medical needs. The work presents simulations to show the advantages of PDC and applies it to compare different black-box XAI tools on a brain MRI dataset for tumor classification, highlighting the importance of rigorous and clinically grounded evaluation frameworks for medical AI

The paper *LLM-Driven Clinical Trial Matching for Lung Cancer Patients: An Explainable Approach* [25] presents MedMatch, a system designed to support oncologists in matching patients to suitable clinical trials. The approach leverages Large Language Models (LLMs) to extract structured clinical parameters from unstructured Electronic Health Records and

to align these features with eligibility criteria of available trials. Unlike commercial cloud-based platforms, MedMatch is deployed on-premises, preserving patient data privacy while providing explainability at multiple levels. The system architecture integrates feature extraction, trial matching, and explanation modules, delivering outputs in structured JSON, criteria-based justifications, and visual evidence highlighting within clinical documents. The paper emphasizes transparency, privacy, and usability, aiming to make AI-driven trial matching more trustworthy and aligned with clinical workflows

2.4. Session 4

The paper *xSTAE: Explaining Sleep Stage Predictions from EEG Signals through Style Transfer Autoencoding* [26] proposes a novel framework for generating counterfactual explanations in time-series classification, with a focus on EEG-based sleep stage prediction. The method, named xSTAE, uses autoencoders to "restyle" misclassified EEG signals into examples of the correct class, revealing the patterns that the classifier failed to detect. This model-agnostic approach allows clinicians and researchers to better understand why errors occur and which signal features are decisive for classification. The work highlights the challenge of explainability for time-series data and contributes an interpretable, generative strategy for uncovering decision-making mechanisms in EEG-based systems, offering a pathway toward more transparent clinical applications

The paper *Vision-Language Models in ECG Interpretation: An Exploratory Study* [27] investigates the use of multimodal AI for cardiovascular diagnostics. Electrocardiograms (ECGs) are essential but complex to interpret, often requiring expert knowledge and time. The authors explore whether Vision-Language Models (VLMs), which combine image and text processing, can provide automated classification and explainable support for ECG interpretation. The study evaluates three models—GPT-4o, PULSE, and a fine-tuned PULSE using Low-Rank Adaptation (LoRA)—on multiple benchmark datasets. The approach involves converting ECG signals into printout-like images and assessing both classification and explanation quality. By focusing on factual accuracy, completeness, and contextual understanding, the work highlights the potential and limitations of applying VLMs to clinical ECG workflows, emphasizing the role of explainability in building trust for medical adoption

The paper *Pilot Assessment of Transparency of LLM-based Systems to Support Emergency Rooms* [28] addresses the challenges of integrating large language models (LLMs) into medical decision support for emergency care. Emergency physicians operate under high stress and time constraints, requiring systems that deliver precise, concise, and trustworthy information. The study explores transparency-related aspects of LLM-based tools, comparing responses from a graph-based retrieval-augmented generation (RAG) system tailored to cardiovascular diseases with outputs from ChatGPT at different temperature settings. The authors describe the development of the MedicalGraphRAG framework, which processes medical literature, builds knowledge graphs, and generates answers with citation support. They further outline a survey of physicians designed to capture preferences for different response styles. The overarching aim is to evaluate how presentation, structure, and transparency of information affect usability and trust in LLM-based decision support within emergency rooms.

The paper *Beyond Static Importance: Quantifying Stability and Distribution Drift* [29] intro-

duces a framework to analyze how feature importance evolves over time in machine learning models, especially in healthcare and time-series contexts. Standard explainability methods like SHAP or LIME provide static insights but fail to capture temporal consistency. The authors propose two complementary metrics: a stability score based on Exponentially Weighted Moving Average (EWMA) to measure how reliably a feature's importance persists, and a distribution drift score based on the Wasserstein distance to track changes in the underlying data distribution. Together, these signals help distinguish shifts due to genuine data dynamics from those caused by model instability. The framework is validated on simulated mental health monitoring data and benchmark time-series datasets, showing its potential to improve trust in AI systems by making explanations temporally robust and interpretable

The paper *Exploring the Expressive Power of Large Language Models in Neuro-Fuzzy System Explainability: A Study on EEG-Based Seizure Detection* [30] investigates how Large Language Models (LLMs) can enhance explanations generated by neuro-fuzzy systems. The study integrates LLMs into a hybrid workflow where fuzzy rules extracted from EEG signals are reformulated into natural language explanations. To evaluate the approach, the authors compare multiple LLM families and sizes, testing different prompting strategies such as zero-shot, persona-based, and fact-checking prompts. The focus is on assessing the linguistic and semantic quality of generated explanations without relying on ground-truth references, which are typically absent in explanation tasks. Applied to seizure detection from EEG data, the framework aims to balance symbolic reasoning with advanced generative models, promoting clearer, more coherent, and user-centered explanations in critical medical contexts

3. Organizing Committee

Gabriella Casalino is currently an Assistant Professor (Tenure Track) at the Computational Intelligence Laboratory (CILab) of the Informatics Department of the University of Bari. Her research is focused on Computational Intelligence methods for interpretable data analysis. She is actively involved in eHealth, Data Stream Mining, and eXplainable Artificial Intelligence. Her work primarily focuses on the medical and educational domains. She holds membership in the IEEE Task Force on Explainable Fuzzy Systems, the Interdepartmental Center for Telemedicine of the University of Bari- CITEL, and the HELMeTO Task Force, which concentrates on Higher Education Learning Methodologies and Technologies Online. She is an active member of the computer science community, contributing to the organization of committees for workshops and special sessions at prestigious international conferences, including ECAI and IEEE WCCI. Additionally, she is an Associate Editor for the international journals Scientific Reports, IEEE Transactions on Computational Social Systems, and Soft Computing. She serves as a Guest Editor for several special issues in the IEEE Transactions on Computational Social Systems and Information Systems Frontiers. She is a Senior member of the IEEE Society and has received

several awards for her research, including the prestigious FUZZ-IEEE Best Paper award.

Giovanna Castellano is an Associate Professor at the Department of Computer Science, University of Bari Aldo Moro, where she coordinates the Computational Intelligence Laboratory (CILab). Her research interests are in the area of Computational Intelligence and Computer Vision. She has been responsible for the local unit of several research projects and is currently the Principal Investigator of the WP 6.4 "Understandability of AI systems" in the NRRP "FAIR - Future Artificial Intelligence Research" project, Spoke 6 - Symbiotic AI. She is an Associate Editor of several international journals. She has been a Guest Editor of special issues and participated in organizing scientific events. She is a reviewer for several international journals published by leading publishers, including Elsevier, IEEE, and Springer, and a member of the program committee of several international conferences. She is a member of the IEEE Society, EUSFLAT Society, INDAM-GNCS Society, IAPR Technical Committee 19 (Computer Vision for Cultural Heritage Applications), CINI-AIIS laboratory, CINI-BIG DATA laboratory, CITEL telemedicine research center, GRIN, MIR laboratories. She is also a member of the IEEE CIS Task Force on Explainable Fuzzy Systems.

Katarzyna Kaczmarek-Majer is an Associate Professor at the Systems Research Institute of the Polish Academy of Sciences. Katarzyna serves as the Principal Investigator for the "Explainable Artificial Intelligence for Monitoring Acoustic Features Extracted from Speech" project (ExplainMe), funded by the European Regional Development Fund. Additionally, she contributes to the "Research of Excellence on Digital Technologies and Wellbeing" project at the Institute for Research and Applications of Fuzzy Modeling, University of Ostrava, Czech Republic. Katarzyna's expertise includes computational intelligence, explainable and trustworthy AI, granular computing, and data stream analysis, with a particular focus on applications in medicine and healthcare. She has co-authored over 50 scientific publications, some of which have been recognised with prestigious awards, such as the Best Paper Award at FUZZ-IEEE 2022 in Padova, Italy, for her work titled "Confidence Path Regularization for Handling Label Uncertainty in Semi-Supervised Learning: A Use Case in Bipolar Disorder Monitoring." She serves on the scientific committees of numerous conferences and is a reviewer for several academic journals and international conferences. Katarzyna is also the Vice-President for the European Society for Fuzzy Logic and Technology and the President of the eHealth Section of the Polish Information Processing Society.

Raffaele Scaringi is a PhD Student at the Computational Intelligence Laboratory (CILab) of the Informatics Department of the University of Bari. He holds a Ph.D. scolarship in Computer Science and Mathematics from the University of Bari, with a dissertation on "Analysis and Enhancement of Artistic Heritage using Artificial Intelligence". His research, conducted within the framework of an Innovative Doctorate with Industrial Connotation, is co-financed by Exprivia S.p.A., focuses on the development of AI-driven methodologies for the analysis, restoration, and valorization of cultural assets. He is also working on different research projects on Explainable Artificial Intelligence and Fuzzy Models using Graph Representation Learning. He is a reviewer for several international journals and conferences published by leading publishers, including Elsevier and Springer.

Gianluca Zaza is an assistant professor at the University of Bari Aldo Moro and is a member of the Computational Intelligence Laboratory (CILab). He is working on "Understandability of AI systems" within the NRRP project "FAIR - Future Artificial Intelligence Research," Spoke 6 - Symbiotic AI. He is the project coordinator for the research project titled "Computational Models based on Fuzzy Logic for eXplainable Artificial Intelligence," which is funded for one year under the "Research Projects GNCS 2023" grant. He is a member of the Interdepartmental Center for Telemedicine of the University of Bari- CITEL. He is a Guest Co-Editor of the Special Issue "Explainability in Human-Centric AI" in Information Systems Frontiers (Springer) and an Associate Editor for the Journal of Intelligent and Fuzzy Systems (IOS Press). He is a senior member of the IEEE Society, and he is also a member of the IEEE CIS Task Force on Explainable Fuzzy Systems. He reviews several international journals published by leading publishers, including Elsevier and Springer.

3.1. Program Committee

- Gabriella Casalino, University of Bari Aldo Moro, Bari, Italy
- Giovanna Castellano, University of Bari Aldo Moro, Bari, Italy
- Marco Cremaschi, University Milano-Bicocca, Milan, Italy
- Paulo Vitor de Campos Souza, Universidade Nova de Lisboa, Lisbon, Portugal
- Pietro Ducange, University of Pisa, Pisa, Italy
- Katarzyna Kaczmarek-Majer, Polish Academy of Sciences, Warsaw, Poland
- Daniel Leite, Paderborn University, Padeborn, Germany
- Corrado Mencar, , University of Bari Aldo Moro, Bari, Italy
- Marcin Ostrowski, Polish Academy of Sciences, Warsaw, Poland
- Daniel Peralta, University of Granada, Granada, Spain
- Marek Reformat, University of Alberta, Alberta, Canada

- Alessandro Renda, University of Trieste, Trieste, Italy
- Souhir Ben Souissi, Bern University of Applied Sciences, Bern, Switzerland
- Jose Sousa, Sano-Centre for Computational Medicine, Kraków, Poland
- Alberto Gaetano Valerio, University of Bari Aldo Moro, Bari, Italy
- Magnus Westerlund, Arcada University of Applied Science, Helsinki, Finland
- Slawomir Zadrozny, Polish Academy of Sciences, Warsaw, Poland
- Gianluca Zaza, University of Bari Aldo Moro, Bari, Italy
- Patryk Żywica, Adam Mickiewicz University, Poznan, Poland

Acknowledgments

The EXPLIMED organizers would like to thank the organizing committee of the 28th European Conference on Artificial Intelligence (ECAI 2028) for hosting this first edition of the workshop. The EXPLIMED workshop was patronized by Fondazione FAIR through the PNRR project, FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007), under the NRRP MUR program funded by NextGenerationEU. Gianluca Zaza and Giovanna Castellano acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGenerationEU. This workshop has been supported by the "INdAM - GNCS Project" CUP E53C24001950001. Ga.C., Gi.C., and G.Z. are members of the CITEL - Centro Interdipartimentale della ricerca in Telemedicina, of the University of Bari Aldo Moro. This work is supported from the project „Research of Excellence on Digital Technologies and Wellbeing CZ.02.01.01/00/22_008/0004583" which is co-financed by the European Union.

Declaration on Generative AI

Generative AI tools, specifically OpenAI's ChatGPT and Grammarly, were used exclusively for grammar correction and language refinement. The authors conceived, wrote, and validated all content.

References

[1] A. Bennetot, I. Donadello, A. El Qadi El Haouari, M. Dragoni, T. Frossard, B. Wagner, A. Sarranti, S. Tulli, M. Trocan, R. Chatila, et al., A practical tutorial on explainable ai techniques, ACM Computing Surveys 57 (2024) 1–44.

[2] M. I. Hossain, G. Zamzmi, P. R. Mouton, M. S. Salekin, Y. Sun, D. Goldgof, Explainable ai for medical data: Current methods, limitations, and future directions, ACM Computing Surveys 57 (2025) 1–46.

[3] L. Nannini, J. M. Alonso-Moral, A. Catalá, M. Lama, S. Barro, Operationalizing explainable artificial intelligence in the european union regulatory ecosystem, IEEE Intelligent Systems 39 (2024) 37–48.

[4] E. H. Houssein, A. M. Gamal, E. M. Younis, E. Mohamed, Explainable artificial intelligence for medical imaging systems using deep learning: a comprehensive review, Cluster Computing 28 (2025) 469.
[5] S. S. Band, A. Yarahmadi, C.-C. Hsu, M. Biyari, M. Sookhak, R. Ameri, I. Dehzangi, A. T. Chronopoulos, H.-W. Liang, Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods, Informatics in Medicine Unlocked (2023). doi:10.1016/j.imu.2023.101286.
[6] F. Di Martino, F. Delmastro, Explainable ai for clinical and remote health applications: a survey on tabular and time series data, Artificial Intelligence Review 56 (2023) 5261–5315.
[7] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. Van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, ACM Computing Surveys 55 (2023) 1–42.
[8] M. A. Mersha, M. G. Yigezu, J. Kalita, Evaluating the effectiveness of xai techniques for encoder-based language models, Knowledge-Based Systems 310 (2025) 113042.
[9] X. Kong, S. Liu, L. Zhu, Toward human-centered xai in practice: A survey, Machine Intelligence Research 21 (2024) 740–770.
[10] J. M. Bauer, M. Michalowski, Human-centered explainability evaluation in clinical decision-making: a critical review of the literature, Journal of the American Medical Informatics Association 32 (2025) 1477–1484.
[11] C. M. van Leersum, C. Maathuis, Human centred explainable ai decision-making in healthcare, Journal of Responsible Technology 21 (2025) 100108.
[12] F. Herrera, Making sense of the unsensible: Reflection, survey, and challenges for xai in large language models toward human-centered ai, arXiv preprint arXiv:2505.20305 (2025).
[13] J. Schneider, Explainable generative ai (genxai): a survey, conceptualization, and research agenda, Artificial Intelligence Review 57 (2024) 289.
[14] X. Zhang, V. S. Sheng, Neuro-symbolic ai: Explainability, challenges, and future trends, arXiv preprint arXiv:2411.04383 (2024).
[15] N. Díaz-Rodríguez, J. Del Ser, M. Coeckelbergh, M. L. de Prado, E. Herrera-Viedma, F. Herrera, Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation, Information Fusion (2023) 101896.
[16] Y. Akaichi, J.-M. Jacquet, I. Linden, W. Vanhoof, Federated inductive logic programming for explainable artificial intelligence, in: Proceedings of the Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy, co-located with 28th European Conference on Artificial Intelligence - ECAI 2025, volume — of *CEUR Workshop Proceedings*, 2025.
[17] L. Bini, M. Sorbi, S. Marchand-Maillet, Massive activations in graph neural networks: Decoding attention for domain-dependent interpretability, in: Proceedings of the Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy, co-located with 28th European Conference on Artificial Intelligence - ECAI 2025, volume — of *CEUR Workshop Proceedings*, 2025.
[18] R. Aasim, G. Zamzmi, J. G. Delfino, J. Jaja, M. A. Lago, Comparative plausibility evaluation of heatmaps for vision transformers in digital mammography, in: Proceedings of the

Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy, co-located with 28th European Conference on Artificial Intelligence - ECAI 2025, volume — of *CEUR Workshop Proceedings*, 2025.

[19] D. Muhammad, M. Salman, M. Bendechache, High cost, low trust? msa-pnet fixes both for medical imaging, in: Proceedings of the Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy, co-located with 28th European Conference on Artificial Intelligence - ECAI 2025, volume — of *CEUR Workshop Proceedings*, 2025.

[20] D. Ramachandram, A. Raval, Crisp-nam: Competing risks interpretable survival prediction with neural additive models, in: Proceedings of the Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy, co-located with 28th European Conference on Artificial Intelligence - ECAI 2025, volume — of *CEUR Workshop Proceedings*, 2025.

[21] J. Karolczak, J. Stefanowski, An interpretable prototype parts-based neural network for medical tabular data, in: Proceedings of the Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy, co-located with 28th European Conference on Artificial Intelligence - ECAI 2025, volume — of *CEUR Workshop Proceedings*, 2025.

[22] I. Heerlien, J. Linssen, L. Gatti, M. Sappelli, B. van Gaal, R. Evering, N. Letwory, Using design thinking for explainable ai: A case study predicting the start of the palliative phase in patients with copd or heart failure, in: Proceedings of the Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy, co-located with 28th European Conference on Artificial Intelligence - ECAI 2025, volume — of *CEUR Workshop Proceedings*, 2025.

[23] N. Babakov, E. Rezgova, E. Reiter, A. Bugarín, Evaluation of explainable ai by medical experts: a survey of the existing approaches, in: Proceedings of the Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy, co-located with 28th European Conference on Artificial Intelligence - ECAI 2025, volume — of *CEUR Workshop Proceedings*, 2025.

[24] N. Blake, D. A. Kelly, S. Calderón Peña, A. Chanchal, H. Chockler, Mrxai: Black-box explainability for image classifiers in a medical setting, in: Proceedings of the Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy, co-located with 28th European Conference on Artificial Intelligence - ECAI 2025, volume — of *CEUR Workshop Proceedings*, 2025.

[25] V. Peppoloni, G. Leone, L. Mazzeo, A. Ferrarin, V. Miskovic, G. Lo Russo, P. Baili, A. Prelaj, F. Corso, Llm-driven clinical trial matching for lung cancer patients: An explainable approach, in: Proceedings of the Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy, co-located with 28th European Conference on Artificial Intelligence - ECAI 2025, volume — of *CEUR Workshop Proceedings*, 2025.

[26] N. Koliou, P. Zazos, C. Romesis, C. Bosch, S. Konstantopoulos, P. Trakadas, xstae: Explaining sleep stage predictions from eeg signals through style transfer autoencoding, in: Proceedings of the Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy, co-located with 28th European Conference

on Artificial Intelligence - ECAI 2025, volume — of *CEUR Workshop Proceedings*, 2025.
[27] S. N. Zeleke, M. Bochicchio, Vision-language models in ecg interpretation: An exploratory study, in: Proceedings of the Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy, co-located with 28th European Conference on Artificial Intelligence - ECAI 2025, volume — of *CEUR Workshop Proceedings*, 2025.
[28] M. Chojnicki, K. Kaczmarek-Majer, P. Burchardt, Y. Ren, M. Z. Reformat, Pilot assessment of transparency of llm-based systems to support emergency rooms, in: Proceedings of the Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy, co-located with 28th European Conference on Artificial Intelligence - ECAI 2025, volume — of *CEUR Workshop Proceedings*, 2025.
[29] M. Ostrowski, O. Hryniewicz, Beyond static importance: Quantifying stability and distribution drift, in: Proceedings of the Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy, co-located with 28th European Conference on Artificial Intelligence - ECAI 2025, volume — of *CEUR Workshop Proceedings*, 2025.
[30] G. Casalino, G. Castellano, D. Margherita, A. G. Valerio, G. Vessio, G. Zaza, Exploring the expressive power of large language models in neuro-fuzzy system explainability: A study on eeg-based seizure detection, in: Proceedings of the Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy, co-located with 28th European Conference on Artificial Intelligence - ECAI 2025, volume — of *CEUR Workshop Proceedings*, 2025.