EXPLAINABILITY 2025

The Second International Conference on Systems Explainability

ISBN: 978-1-68558-318-7

October 26th - 30th, 2025

Barcelona, Spain

EXPLAINABILITY 2025 Editors

Ray Jones, University of Plymouth, UK

EXPLAINABILITY 2025

Forward

The Second International Conference on Systems Explainability (EXPLAINABILITY 2025), held on October 26-30, 2025 in Barcelona, Spain, continued a series of events dealing with models and metrics to build a documented and provable trust for the developers and users of any kind of system. Explainability helps to validate tracking between system design requirements and current implementation ensuring validation of evolving properties by continuously learning and adapting the original requirements.

Interpretability, Explainability, and Understandability are characteristics needed for any product, system, device, government regulation, or societal law to increase their trustfulness and acceptability by the end-users. Their role is to avoid bias and increase confidence in the systems' output.

Explainability favors interpretability and understandability and should be considered during the requirements, design, deployment and maintenance phases of all software, hardware, and complex systems. To a large extent, explainability is present as a user manual, software requirements tracking and code identification, validation/testing results, interactive interfaces, explanation of models, guidelines for industrial robots, and in any human-driven procedural processes. Desiderata on explainability become more complex for Artificial Intelligence (AI)-based entities/systems in terms of 'thinking' via internal mechanisms and accepting/trusting the output.

Explainability is a sought-after property of any complex 'products'. In AI-based systems, the explanation of the behavior of models for certain critical systems is mandatory. This is a complex task, considering that the behavior is the result of intricate development processes involving humans, algorithms, datasets, and other artificial entities (tools).

This conference was very competitive in its selection process and very well perceived by the international community. As such, it attracted excellent contributions and active participation from all over the world. We were very pleased to receive a large amount of top quality contributions.

We take here the opportunity to warmly thank all the members of the EXPLAINABILITY 2025 technical program committee as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and efforts to contribute to the EXPLAINABILITY 2025. We truly believe that thanks to all these efforts, the final conference program consists of top quality contributions.

This event could also not have been a reality without the support of many individuals, organizations and sponsors. We also gratefully thank the members of the EXPLAINABILITY 2025 organizing committee for their help in handling the logistics and for their work that is making this professional meeting a success.

We hope the EXPLAINABILITY 2025 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in system explainability research. We also hope that Barcelona provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city

EXPLAINABILITY 2025 Steering Committee

Thomas Fehlmann, Euro Project Office AG, Zurich, Switzerland
Mahdi Jalili, RMIT University, Australia

Fairouz D Kamareddine, Heriot-Watt University, Scotland
Ray Jones, University of Plymouth, UK

EXPLAINABILITY 2025

Committee

EXPLAINABILITY 2025 Steering Committee

Thomas Fehlmann, Euro Project Office AG, Zurich, Switzerland
Mahdi Jalili, RMIT University, Australia
Fairouz D Kamareddine, Heriot-Watt University, Scotland
Ray Jones, University of Plymouth, UK

EXPLAINABILITY 2025 Technical Program Committee

André Artelt, Bielefeld University, Germany
Esteban Bautista, Université du Littoral Côte d'Opale, France
Dorota Bielińska-Wąż, Medical University of Gdańsk, Poland
Dalmo Cirne, Workday Inc., USA
Thomas Fehlmann, Euro Project Office AG, Zurich, Switzerland
Ming Gong, Bing Experiences | Microsoft, China
Nada Ahmed Hamed Sharaf, German International University, Cairo, Egypt
Amr Hendy, Microsoft, USA
Mahdi Jalili, RMIT University, Australia
Fairouz Kamareddine, Heriot-Watt University, Edinburgh, UK
John Kos, Georgia Institute of Technology, USA
Shudong Liu, University of Macau, Macau
Giovanni Montana, University of Warwick, UK
Minheng Ni, Hong Kong Polytechnic University, Hong Kong
Albert Solé Ribalta, Universitat Oberta de Catalunya, Spain
Swati Tyagi, JP Morgan Chase & Co., Wilmington, DE, USA
Piotr Wąż, Medical University of Gdańsk, Poland
Chang D. Yoo, Korea Advanced Institute of Science and Technology (KAIST), South Korea
Abdou Youssef, George Washington University, USA
Daqing Yun, Harrisburg University, USA

Copyright Information

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

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

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

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

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

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

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

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

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

Table of Contents

Identifying Confusion Trends in Concept-based XAI for Multi-Label Classification

Haadia Amjad
Chair of Fundamentals of Electrical Engineering
TUD Dresden University of Technology
Dresden, Germany
e-mail: haadia.amjad@tu-dresden.de

Kilian Göller
Chair of Fundamentals of Electrical Engineering
TUD Dresden University of Technology
Dresden, Germany
e-mail: kilian.goeller@tu-dresden.de

Steffen Seitz
Chair of Fundamentals of Electrical Engineering
TUD Dresden University of Technology
Dresden, Germany
e-mail: steffen.seitz@tu-dresden.de

Carsten Knoll
Chair of Fundamentals of Electrical Engineering
TUD Dresden University of Technology
Dresden, Germany
e-mail: carsten.knoll@tu-dresden.de

Ronald Tetzlaff
Chair of Fundamentals of Electrical Engineering
TUD Dresden University of Technology
Dresden, Germany
e-mail: ronald.tetzlaff@tu-dresden.de

***Abstract*—Deep Neural Networks (DNNs) deployed in high-risk domains, such as healthcare and autonomous driving, must be not only accurate but also understandable to ensure user trust. In real-world computer vision tasks, these models often operate on complex images containing background noise and are heavily annotated. To make such models explainable, Concept-based Explainable AI (CXAI) methods need to be assessed for their applicability and problem-solving capacity. In this work, we explore CXAI use cases in multi-label classification by training two DNNs, VGG16 and ResNet50, on the 20 most annotated labels in the MS-COCO dataset (Microsoft Common Objects in Context). We apply two CXAI methods, CRP (Concept Relevance Propagation) and CRAFT (Concept Recursive Activation FacTorization), to generate concept-level explanations and investigate the overall evaluations. Our analysis reveals three key findings: (1) CXAI highlights learning weaknesses in DNNs, (2) higher concept distinctiveness reduces label and concept confusion, and (3) environmental concepts expose dataset-induced biases. Our results demonstrate the potential of CXAI to enhance the understanding of model generalizability and to diagnose bias instigated by the dataset.**

Keywords-Concept-based XAI; Multi-Label Classification; Concept Distinctiveness.

I. INTRODUCTION

Deep Neural Network (DNN) [1] performance is crucial for their adoption in real-world applications. However, understanding their decisions is also important, especially in high-risk domains like autonomous driving and medical diagnosis. Real-world datasets often vary in resolution and object size, with complex scenes including small, clustered, or overlapping objects. Multi-label datasets, where images have multiple annotations, frequently suffer from class imbalance. This can lead to confusion (i.e., errors made in predicting the correct class/data points) between labels and wrong associations. Even high-performing models that exhibit confusion need deeper analysis. Explainable AI (XAI) methods are useful in revealing these learning patterns [2].

XAI provides interpretability for black-box models [2]. Concept-based XAI (CXAI) identifies semantically meaningful features relevant to a class [3], unlike saliency maps, which are harder to interpret in complex scenes [4]. Concepts reflect how a DNN internally represents a class [5]. However, DNNs may learn unintended associations, concept bias or spurious correlations, where background elements influence classification (e.g., associating "fingers" with a pen) [6]. We refer to non-target concepts produced by such bias as "environmental concepts."

CXAI methods often visualize activation maps or focused image regions [7]. These show both target and environmental concepts. Determining whether an environmental concept is valid requires further analysis. Its presence may reflect dataset bias or mislearning.

In this work, we train two state-of-the-art DNNs, ResNet50 and VGG-16, on the 20 most annotated MS-COCO labels [8]. Using two model checkpoints per architecture, one well-performing and one poor, we evaluate their predictions using CXAI methods: CRP [9] and CRAFT [10]. These methods produce focused region

 ISBN: 978-1-68558-318-7

visualizations and scores that determine a concept's contribution to the overall learning (concept importance) or target label learning (concept relevance) of the DNN model. We compare results using concept error and distinctiveness (see Section III, D) to study confusion trends across models.

The main contributions of this paper can be summarized as follows:

- We demonstrate that CXAI methods can reveal learning weaknesses in deep neural networks.
- We find that greater concept distinctiveness is associated with reduced confusion in label predictions and concept attributions.
- We show that environmental concepts can expose dataset-induced biases in model learning and interpretation.

The remainder of this paper is organized as follows: Section II reviews related studies. Section III describes the experimental setup, including the dataset, DNN models, CXAI methods, and key terminology. Section IV presents the results structured around our three main contributions. Finally, Section V concludes the paper and discusses directions for future work.

II. Related Work

Various CXAI methods are available for use today, and it is a growing research field. Lee et al. [11] detail the current state of CXAI methods. Their study identifies three main directions for future research: the choice of concepts to explain, the selection of concept representation, and methods to control concepts.

Some studies focus on using concepts to detect potential biases in DNN models. Their evaluation emphasizes the relationship between different concepts and classes and aims to expose potential biases in the learning of the DNN. Singh et al. [12] study model biases in both, the model learning process and the model's semantic understanding (concept biases), by evaluating the DNN model's ability to recognize a class in the presence and absence of the established context (via learning) for a multi-label classification task.

With newer emerging methods in the realm of CXAI, the desire to fully understand how they can be effectively used with AI systems increases. The dataset, for example, is an important factor contributing to the meaningfulness of the explainability method. The evaluation of CXAI by Ramaswamy et al. [13] addresses important considerations for CXAI methods that influence their effective usage. They emphasize that the impact of the choice of the dataset, even with slight variations in the dataset options, changes the model decision and the explanation provided by a CXAI method.

To study the relationship between confusion and concept-based explanations, we select two CXAI methods to answer the "where" (*..the important information is*) and "what" (*..is the important information*) questions. CRP, proposed by Achtibat et al. [9], is based on the Layer-wise Relevance Propagation (LRP) method [14]. CRP addresses "what" and "where" explanations by exploiting concepts in hidden layers of a DNN model and locating them in the input data. It assesses the contribution of each concept for a target class; in other words, it introduces concept relevance. CRP utilizes relevance maximization to tune its visualization, which depicts a series of focused concepts. CRAFT is another "what" and "where" method proposed by Fel et al. [10] based on the Grad-CAM method [15]. They utilize Sobol indices to estimate the importance of concepts that have been identified using Non-Negative Matrix Factorization (NMF) recursively, generating sub-concepts (concepts of smaller, more focused areas in the image).

Existing research has advanced CXAI by defining concepts and applying them to detect biases and assess dataset effects. Building on this foundation, our work investigates how confusion interacts with concept-based explanations through the lens of CRP and CRAFT.

III. Experimental Setup

Just as with any other explainable AI pipeline, our experimentation contains the training of DNNs model and its evaluations and the usage of an XAI method and its evaluation, illustrated in Figure 1. This section contains details of our workflow.

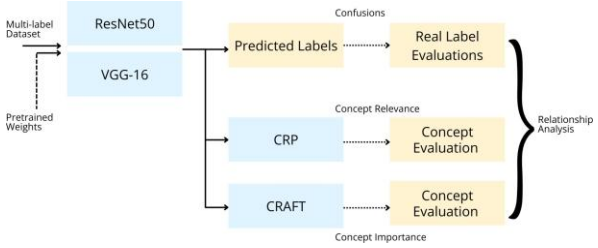

Figure 1. Schematic diagram of our experimental setup.

A. Dataset

MS-COCO [8] is a large-scale dataset widely used for computer vision tasks such as object detection, captioning, segmentation and classification. The 2017 object detection subset includes 80 "things" classes, objects with clear boundaries, across 118,000 images. As test labels are unavailable, we split the training set 90/10, resulting in 106,200 training and 11,800 test images. For our experiments, we focus on the 20 most frequently annotated labels in the training set to ensure sufficient data per class and meaningful inter label relationships.

B. DNN Models

ResNet50 [16] is a popular image classification model due to its residual learning feature, which mitigates information loss. It balances accuracy and efficiency well, and its ImageNet-pretrained weights are widely used [17].

VGG-16 [17], known for its simple and uniform structure of stacked convolutional and fully connected layers, is often used as a baseline for deep learning applications. Despite its larger parameter count, it performs well on classification tasks and is easy to implement.

These two models are chosen as they are widely used, and many XAI methods have been proven to work with them. Some of the latest models require large adaptations of XAI methods to be made [18]. Our study focuses on base-level use cases, to be adaptable across different domains; hence, we train ResNet50 and VGG-16 models, pretrained on ImageNetV2 [19], using PyTorch for 350 epochs, saving all checkpoints. For each model, we select two "scenarios" from the saved checkpoints:

- **Scenario 1 (well-performing model):**
  - **ResNet50:** Accuracy: 82.85%, Recall: 85.50, Precision: 58.84, F1 Score: 60.84
  - **VGG-16:** Accuracy: 84.26%, Recall: 86.91, Precision: 59.74, F1 Score: 58.84
- **Scenario 2 (poor-performing model):**
  - **ResNet50:** Accuracy: 58.24%, Recall: 77.04, Precision: 53.82, F1 Score: 42.92
  - **VGG-16:** Accuracy: 52.85%, Recall: 74.50, Precision: 53.62, F1 Score: 46.12

These scenarios are created to have two different sets of performance metrics against which to evaluate explainability. We evaluate models using accuracy, recall, precision, F1 score, and confusion matrices tailored for multi-label tasks. Specifically, we use the multi-label confusion tensor by Krstinic´ et al. [20], which accounts for label imbalance—well-suited for the MS-COCO dataset.

We also compute Mutual Information (MI) and Jaccard Similarity Coefficient (JSC) between labels. We use these metrics to understand which target labels are more likely to share information or similarities with which predicted labels.

C. CXAI Methods

We investigate the effect of confusion on two CXAI methods, CRAFT and CRP, across all four model scenarios.

- **CRAFT** outputs concept *importance*, representing the overall contribution of each concept to the model's learning process.
- **CRP** provides concept *relevance*, indicating the contribution of a concept to specific target classes.

While both methods offer different perspectives, we do not compare them directly or suggest one is superior. Instead, we use their outputs to explore how label confusion is reflected in learned concepts.

We compute concept distinctiveness [21] and concept error [22] for both methods. Concept error is evaluated against a subjective ground truth (detailed in the next section). Additionally, we adapt mutual information to measure shared information between concepts and compare these findings to our DNN evaluations to support our hypotheses.

D. Explanation of terms (in brief)

This sub-section briefly explains some terminologies in CXAI and our adaptations.

1) Concept Distinctiveness: Concept distinctiveness, defined in Eq. (1), measures how unique a concept is compared to others, with values ranging from 0 to 1. Low distinctiveness suggests overlapping or redundant concepts, which may indicate learning errors [21].

$$D(C_i, C_j) = 1 - \frac{v_{C_i} \cdot v_{C_j}}{|v_{C_i}||v_{C_j}|} \quad (1)$$

Here, v_{C_i} and v_{C_j} are the concept vectors for concepts C_i and C_j, respectively. Concept vectors are directions in activation space that capture distinct features [23].

2) Concept Error: Concept error captures incorrect or irrelevant concept usage during prediction [22]. To approximate accuracy (in binary classification), we define a rough "ground truth" by selecting only those concepts that belong to the target class, excluding environmental concepts. This approach offers an estimate of model confusion, though a structured human study is recommended for practical validation.

3) Mutual Information: Mutual information (MI) quantifies the dependency between two variables. In multi-label classification, it measures how much information one label provides about another. Applied to concepts, MI reflects how much information is shared between two concept vectors, revealing potential dependencies or redundancies in learned features [24].

IV. Results

In this section, we present our findings based on case studies of different label evaluations. These case studies comprise comparisons of the evaluations described in the previous section.

A. Confusion in Labels Can Be Understood by Their Explanations

TABLE I. TOP CONFUSION AND MUTUAL INFORMATION SCORES IN SCENARIO 1 OF RESNET50

Class Name	Top Confusion Class				Top MI Class				Jaccard Similarity
	1st	Score	2nd	Score	1st	MI Score	2nd	MI Score	
person	car	1148.00	chair	1081.70	handbag	0.0221	backpack	0.0176	0.6008
car	truck	207.17	bench	173.16	truck	0.0400	traffic light	0.0280	0.1894
motorcycle	truck	86.33	handbag	85.70	car	0.0094	person	0.0037	0.1474
truck	airplane	118.35	car	117.22	car	0.0399	boat	0.0020	0.0077
boat	Parking meter	89.70	car	76.30	chair	0.0682	fork	0.0017	0.0068

TABLE II. TOP CONFUSION AND MUTUAL INFORMATION SCORES IN SCENARIO 2 OF RESNET50

Class Name	Top Confusion Class				Top MI Class				Jaccard Similarity
	1st	Score	2nd	Score	1st	MI Score	2nd	MI Score	
person	backpack	1995.20	bench	1922.50	tie	0.0221	umbrella	0.0176	0.5470
car	backpack	340.70	bench	334.80	boat	0.0399	stop sign	0.0280	0.1159
motorcycle	backpack	277.60	handbag	273.41	bicycle	0.0372	car	0.0199	0.1289
truck	backpack	209.07	bench	200.09	motorcycle	0.0399	Fire hydrant	0.0077	0.0755
boat	car	134.04	bird	133.66	fork	0.0017	refrigerator	0.0010	0.0440

TABLE III. PERCENTAGE OF CO-OCCURRENCE OF TARGET LABEL WITH OTHER LABELS (TOP 20 FREQUENTLY ANNOTATED LABELS)

Class Name	1st Class	%	2nd Class	%
person	car	13.29	backpack	7.85
car	person	69.54	backpack	8.43
motorcy -cle	person	79.55	car	39.32
truck	person	65.15	car	59.80
boat	person	65.69	car	8.66
traffic light	car	61.22	person	59.19
bench	person	73.75	car	14.63
bird	person	24.56	boat	7.29
sheep	person	24.07	dog	7.59
backpack	person	91.06	car	18.69
umbrella	person	86.87	handbag	28.81
handbag	person	90.95	backpack	24.62
kite	person	92.84	car	11.54
bottle	person	53.65	cup	34.65
cup	person	52.76	dining table	50.92
bowl	dining table	47.76	person	40.73
banana	person	41.37	bowl	23.05
potted plant	person	44.07	chair	38.61
dining table	person	49.58	chair	43.29
book	dining table	75.61	cup	52.97

TABLE IV. CXAI METHOD EVALUATION COMPARED WITH CONFUSION SCORE FOR 'PERSON' LABEL

Label	Model	CXAI Method	Concept Error	Concept Distinct -iveness	Confus -ion Score
Person	ResNet Scenario 1	Craft	0.20	0.76	0.09
Person	ResNet Scenario 2	Craft	0.38	0.48	0.26
Person	VGG-16 Scenario 1	Craft	0.24	0.71	0.12
Person	VGG-16 Scenario 2	Craft	0.41	0.43	0.28

Label confusion occurs when models struggle to distinguish between classes with overlapping features or co-occurring contexts, often due to ambiguous data, mislabeling, or internal misinterpretation. We hypothesize that CXAI methods, particularly through MI and concept distinctiveness, can reveal whether confusion stems from visual similarity, dataset bias, or how the model encodes relationships between labels.

Tables I and II present confusion and MI scores for three highly confused classes across both ResNet50 scenarios. In scenario 1, *person* is confused with *car* and *chair*, while *car* overlaps with *truck* and *bench*. MI analysis shows that *person* shares high information content with *handbag* and *backpack*, and *car* with *truck* and *traffic light*. These associations indicate that the model is not learning isolated class-specific features, but instead forming dependencies based on recurring visual or contextual co-occurrence. Table III supports this, showing frequent joint appearance of labels such as *person* and *accessories*, or *car* and *truck*, which reinforces these spurious links.

Table IV further highlights the role of CXAI metrics

in understanding confusion. In scenario 1, where models perform better, *person* has lower concept error and higher distinctiveness, aligning with reduced confusion. In scenario 2, we observe the opposite: increased concept error, lower distinctiveness, and significantly higher confusion scores. These patterns suggest that when a model lacks distinct conceptual boundaries between classes, it tends to rely more heavily on misleading contextual aspects.

Together, these findings show how CXAI methods help expose the roots of confusion. By combining explanations with performance metrics and co-occurrence statistics, we gain a clearer view of when confusion reflects real-world visual similarity versus when it results from dataset bias or poor internal representations.

B. Distinctiveness Reduces Conceptual Confusion

When a concept is distinct, its features are unique and specific, allowing it to be more accurately defined and recognized. In contrast, concepts derived from confused or overlapping labels tend to be "confused" themselves, as they learn features that are shared across multiple classes rather than those unique to their true class. This issue arises from concept bias, where the model may associate a class with irrelevant features that co-occur with other classes, as shown in Figure 2.

Figure 2. Concepts of class "tennis racket" in scenario 1 of VGG-16. We can see that "person" is heavily present in these explanations.

TABLE V. MUTUAL INFORMATION, CONCEPT DISTINCTIVENESS, AND CONCEPT ERROR IN SCENARIO 1 OF RESNET50

Class Name	Top MI (Concept)		Lowest Distinctive (CRP)		Lowest Distinctive (CRAFT)		Concept Error
	1st	2nd	1st	2nd	1st	2nd	Value
person	car	backpack	car	backpack	car	tennis racket	0.7291
car	truck	bus	truck	traffic light	handbag	truck	0.5385
dining table	chair	cup	chair	fork	person	chair	0.0166

From the information given in Table VI, it is evident that a poor-performing model is not ideal for concept-based explanations due to the lack of clear distinctions between classes. This can be seen in scenario 2 of ResNet50, where classes like person show less distinctiveness with other unrelated classes. In scenario 1, shown in Table V, we see a more effective distinction between highly confused classes like car and person, which

TABLE VI. MUTUAL INFORMATION, CONCEPT DISTINCTIVENESS, AND CONCEPT ERROR IN SCENARIO 2 OF RESNET50

Class Name	Top MI (Concept)		Lowest Distinctive (CRP)		Lowest Distinctive (CRAFT)		Concept Error
	1st	2nd	1st	2nd	1st	2nd	Value
person	car	tennis racket	backpack	bottle	backpack	umbrella	0.8136
car	truck	traffic light	bench	fire hydrant	backpack	boat	0.6388
dining table	cup	bottle	chair	fork	person	potted plant	0.0753

indicates that a well-performing model actively tries to separate these difficult-to-distinguish classes (previously established based on confusion scores, see Table I, V, VI and III).

By focusing on distinctiveness metrics and correlating them with confusion patterns in Table I and co-occurrence in Table III, we see that increasing concept distinctiveness can significantly aid in or point to improved model performance. This insight not only helps in diagnosing where models are struggling but also guides how to curate datasets and improve feature learning to reduce confusion and improve overall classification accuracy.

C. Environmental Concepts Reveal Dataset Biases

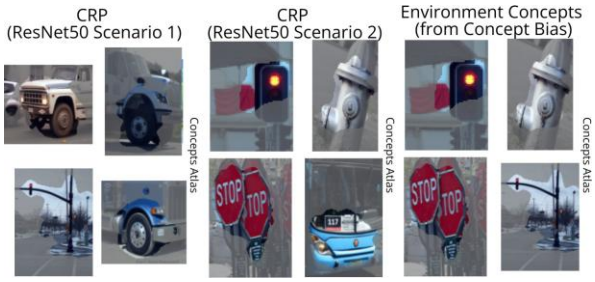

Figure 3. Environmental concepts generated from CRP for class "car" in scenario 1 and 2 of ResNet50.

TABLE VII. MUTUAL INFORMATION (CONCEPT), MUTUAL INFORMATION AND CONFUSION SCORES IN SCENARIO 1 OF VGG-16

Class Name	Top MI (Concept)		Top M (Class)		Top Co fusion	
	1st	2nd	1st	2nd	1st	2nd
umbrella	person	handbag	backpack	handbag	person	car
dining table	chair	fork	chair	cup	apple	person
traffic light	person	car	car	fire hydrant	person	car

Environmental concepts emerge from concept bias and often reflect patterns in the training dataset. We observe that classes within the same "supercategory" (e.g., *sports*: *baseball glove*, *tennis racket*) tend to produce biased explanations, frequently including environmental concepts from related classes, illustrated in Figure 3. This suggests that, beyond model performance, the diversity and distinctiveness of training samples play a key role in learning meaningful class representations.

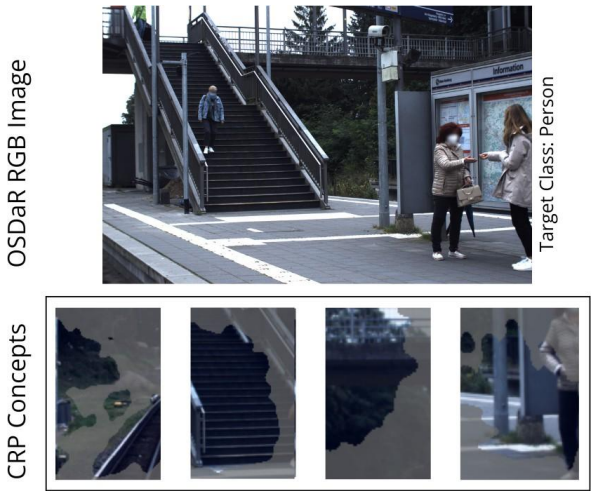

Figure 4. Concepts generated by CRP on OSDaR23 dataset for class "person".

Table VII illustrates the top mutual information and confusion scores for selected classes. For instance, *dining table* in scenario 1 is frequently associated with *chair*, *person*, *apple*, and *cup*, which are labels that share semantic but not structural similarity. Such associations, while intuitive to humans, suggest that the model is not generalizing but instead relying on frequent co-occurrences, which is problematic in deployed systems. High concept error rates for classes like *umbrella*, *person*, *handbag*, and *car*, paired with low distinctiveness scores between semantically unrelated objects (e.g., *umbrella* and *traffic light*), reinforce this concern, especially when models perform poorly.

To further support this, we evaluate OSDaR23 [25], a multi-sensor dataset for autonomous train driving. Despite strong accuracy (95.92%) and F1 (79.93) on a ResNet50 model trained on its RGB subset, CXAI explanations reveal low generalizability. Since *person* consistently appears near platforms or staircases, CRP visualizations heavily rely on these backgrounds, none of which are labeled in the dataset, as illustrated in Figure 4. As a result, *person* has the lowest distinctiveness score with *track*, and a high concept error, indicating dangerous misattribution.

These findings highlight how environmental concepts reveal dataset-induced biases that compromise generalization. In real-world or high-risk applications, such as autonomous systems, these misleading correlations can reduce model reliability. Diverse and well-annotated datasets are essential to prevent concept bias and ensure models learn robust, semantically accurate representations.

V. Conclusion and Future Work

Our study demonstrates that confusion in multi-label classification is directly reflected in concept-based explanations. By comparing model evaluations with CXAI properties, we observe that label confusion often results from overlapping or spurious environmental concepts, emphasizing the role of CXAI in uncovering learning biases and assessing model generalizability. We further show that concept distinctiveness is inversely related to conceptual confusion, models with higher distinctiveness show clearer feature boundaries and reduced bias, while lower distinctiveness leads to shared or incorrect associations across classes. CRP and CRAFT help identify such conceptual ambiguities, making them useful tools for model diagnosis. Finally, our results highlight that environmental concepts can reveal dataset-induced biases, especially in cases where co-occurring objects affect model learning. In datasets with label imbalance or strong contextual patterns, models may form misleading correlations, reducing their ability to generalize. This is particularly problematic in high-risk applications, reinforcing the need for diverse, well-annotated datasets to ensure robust and reliable AI models. For future work, this case study can be extended to more complex models and datasets.

Acknowledgement

This work is partly supported by BMFTR (Federal Ministry of Research, Technology and Space) in DAAD project 57616814 (SECAI, School of Embedded Composite AI, https://secai.org/).

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, 2015.

[2] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, “Explaining deep neural networks and beyond: A review of methods and applications,” *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, 2021.

[3] E. Poeta, G. Ciravegna, E. Pastor, T. Cerquitelli, and E. Baralis, *Concept-based explainable artificial intelligence: A survey*, Preprint arXiv:2312.12936, 2023.

[4] R. Müller, M. Thoß, J. Ullrich, S. Seitz, and C. Knoll, “Interpretability is in the eye of the beholder: Human versus artificial classification of image segments generated by humans versus xai,” *International Journal of Human-Computer Interaction*, pp. 2371–2393, 2024.

[5] S. S. Y. Kim, E. A. Watkins, O. Russakovsky, R. Fong, and A. Monroy-Hernández, “"help me help the ai": Understanding how explainability can support human-ai interaction,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–17.

[6] P. Stock and M. Cisse, “Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 504–519.

[7] B. Kim et al., "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International Conference on Machine Learning*, PMLR, 2018, pp. 2668–2677.

[8] T.-Y. Lin et al., "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, Springer, 2014, pp. 740–755.

[9] R. Achtibat et al., "From attribution maps to human-understandable explanations through concept relevance propagation," *Nature Machine Intelligence*, vol. 5, pp. 1006–1019, 2023.

[10] T. Fel et al., "Craft: Concept recursive activation factorization for explainability," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2711–2721.

[11] J. H. Lee, G. Mikriukov, G. Schwalbe, S. Wermter, and D. Wolter, "Concept-based explanations in computer vision: Where are we and where could we go?" In *European Conference on Computer Vision*, Springer, 2025, pp. 266–287.

[12] K. K. Singh et al., "Don't judge an object by its context: Learning to overcome contextual bias," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 070–11 078.

[13] V. V. Ramaswamy, S. S. Y. Kim, R. Fong, and O. Russakovsky, "Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 932–10 941.

[14] A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, and W. Samek, "Layer-wise relevance propagation for neural networks with local renormalization layers," in *International Conference on Artificial Neural Networks*, Springer International Publishing, 2016, pp. 63–71.

[15] R. R. Selvaraju et al., "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] H. Ibrahim Aysel, X. Cai, and A. Prugel-Bennett, "Explainable artificial intelligence: Advancements and limitations," *Applied Sciences*, vol. 15, p. 7261, 2025.

[19] J. Deng et al., "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2009, pp. 248–255.

[20] D. Krstinic´, A. K. Skelin, I. Slapnicˇar, and M. Braovic´, "Multi-label confusion tensor," *IEEE Access*, vol. 12, pp. 9860–9870, 2024.

[21] B. Wang, L. Li, Y. Nakashima, and H. Nagahara, "Learning bottleneck concepts in image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 962–10 971.

[22] P. W. Koh et al., "Concept bottleneck models," in *International Conference on Machine Learning*, PMLR, 2020, pp. 5338–5348.

[23] L. O'Mahony, V. Andrearczyk, H. Müller, and M. Graziani, "Disentangling neuron representations with concept vectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3770–3775.

[24] M. E. Zarlenga et al., *Concept embedding models: Beyond the accuracy-explainability trade-off*, Preprint arXiv:2209.09056, 2022.

[25] R. Tagiew et al., "Osdar23: Open sensor data for rail 2023," in *International Conference on Robotics and Automation Engineering*, 2023, pp. 270–276.

CAT in the Box: A CausalAI – Tsetlin Machine Duo Enabling explainable Stroke Diagnosis and Prevention

Jalpa Soni, Emelian Gurei, Jaime Lopez Sahuquillo, Sergio García Gomez, Victor M. Saenger
AI Innovation Lab, Capitole Consulting,
Balmes, 89, 08008 - Barcelona, Spain
email: jalpabensoni@capitole-consulting.com

Manuel Rodriguez Yañez, Francisco Campos Perez
Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS),
Clinical Hospital, D Building, 1st Floor,
Travesía da Choupana S/N, 15706 Santiago de Compostela, Spain

***Abstract*— In this paper, we propose an explainable framework to assess biomarker significance in brain stroke data by combining Causal Artificial Intelligence (AI), which models cause–effect relationships beyond simple correlations, with a Tsetlin Machine, a symbolic rule-based learning algorithm that generates human-readable logic clauses. In a first step, Causal AI is used to uncover complex interdependencies among biomarkers and to identify the most impactful ones, while the interpretable clauses of the Tsetlin Machine enhance understanding and support improved diagnosis, prognosis, and prevention in stroke patients. This methodological strategy sets a novel foundation for better understanding of complex brain diseases.**

Keywords - Brain stroke; Causal AI; Explainability; Interpretability; Tsetlin Machine.

I. INTRODUCTION

Stroke, caused by an alteration of the blood supply to the brain, is a medical emergency that requires immediate attention in urgent care departments and specialized stroke units. It is a leading cause of long-term disability and the second leading cause of death globally. In Spain, about 1 in 5 stroke patients are readmitted with a recurrent stroke [1][2]. These statistics highlight the importance of early and accurate diagnosis, as timely intervention can significantly reduce mortality and long-term disability. Despite notable advances in medical imaging and diagnostics, deciphering the intricate relationships among stroke-related biomarkers remains a significant challenge.

In recent years, Machine Learning (ML) has shown promise for detecting subtle patterns in biomedical data [3]. However, many ML models lack transparency, offering limited insight into how predictions are made. This opacity poses a major barrier to their adoption in clinical settings, where trust, accountability, and explainability are essential for informed decision-making.

In this paper, we propose a novel approach that integrates Causal AI [4] to model cause-effect relationships rather than simple correlations among stroke-related biomarkers with Tsetlin Machines [5][6][8][9], a symbolic, rule-based learning model that can uncover and help interpret how specific biomarkers influence stroke outcomes. Causal AI refers to machine learning methods that model cause–effect relationships, beyond mere correlations, whereas Tsetlin Machines are interpretable, rule-based learning models that construct human-readable logic clauses for classification tasks [6]. For example, a Tsetlin Machine might generate a rule such as: *"If LDL cholesterol is high **and** age is above 65, **and** prior use of antiplatelet drugs is absent, then the patient is more likely to suffer an ischemic stroke."* Such clauses are easily understandable by clinicians and can be directly compared with established medical knowledge. Together, these not only enhance predictive accuracy, but also provide a transparent, interpretable insight essential for clinical decision-making.

The rest of the paper is organized as follows. In **Section II**, we describe the methodology, including an overview of the dataset, pre-processing steps, the application of Causal AI, and the use of Tsetlin Machines for interpretable classification. In **Section III**, we present and discuss the results obtained from both the causal inference analysis and the Tsetlin Machine model, highlighting their clinical relevance. In **Section IV**, we conclude the paper by summarizing the key findings and outlining directions for future research and model improvements.

II. METHODOLOGY

In this section, we describe the methodology, with subsections on an overview of the dataset, pre-processing steps, Causal AI, and the Tsetlin Machines.

A. Overview

As mentioned in the introduction, we employ a hybrid methodology that combines Causal AI, a set of techniques designed to model cause–effect relationships rather than mere correlations, with Tsetlin Machines, symbolic rule-based learning algorithms capable of generating human-readable logic clauses. This integrated approach allows us to both identify the underlying causal relationships among

biomarkers that drive clinical outcomes in stroke diagnosis and prognosis, and to extract interpretable rules that clarify how specific biomarker patterns contribute to different stroke subtypes. By linking causal discovery with transparent classification, our method not only improves predictive power but also enhances clinical trust and explainability. The study has received the ethical approval of the Santiago/Lugo clinical ethical committee (code: 2025/221).

B. Dataset and pre-processing

The dataset consists of about 4000 data points with 62 features, containing relevant clinical, demographic and biochemical biomarkers. Standard pre-processing steps were applied, as listed below:

- Removal of non-relevant features using domain knowledge (e.g., multiple stroke determination tests at various times would dominate causal relations, suppressing the weight of other biomarkers).
- Missing value imputation using binary and iterative imputers, which estimate missing values by iteratively predicting them based on other available features. This is particularly useful in this data set as the relationships between medical features can provide valuable information for filling in missing data. This is done for binary and non-binary features respectively.

C. Causal AI

To identify potential causal relationships among biomarkers, we applied the PC algorithm (after its authors, Peter and Clark), a constraint-based causal discovery method, to the pre-processed dataset [7]. At this stage, the dataset contains approximately 50 features including the target (type of stroke – ischemic or haemorrhagic).

Since our objective is to isolate the most influential biomarkers, we employed two graph-theoretic measures to rank nodes (features) within the causal graph:

- *Degree Centrality:* Measures the number of direct connections for a node. High degree centrality suggests that a feature has broad influence.
- *Betweenness Centrality:* Quantifies how often a node appears on the shortest paths between other nodes. High betweenness centrality implies that a feature is a critical intermediary or bridge in the causal network.

To minimize selection bias to ensure that both direct and indirect influences are taken into account, we first created two separate ranked lists of features: one based on degree centrality and the other based on betweenness centrality. From each ranking, we extracted the top 25 features, representing those with the strongest influence according to the respective measures. Next, we introduced a composite centrality score, which assigns weights to features depending on their positions in the two rankings, thereby balancing the contribution of both centrality measures. Finally, by comparing the two lists and focusing on the features with the highest combined scores, we identified the 10 most influential biomarkers that consistently appeared as important across both centrality perspectives.

D. Tsetlin Machines

Following the identification of the top 10 biomarkers through causal inference, we applied a rule-based convergence Tsetlin Machine (TM) [8][9][10] to model their relationship with stroke subtypes. This model is a logic-based learning algorithm that constructs human-interpretable propositional logic clauses to perform classification. It operates by learning patterns expressed as conjunctive logical clauses, where each clause is essentially a combination of conditions that must be satisfied for a prediction to be made (for example, *if biomarker A is present and biomarker B is absent, then the case belongs to class X*). Rather than relying on a single clause, the Tsetlin Machine generates a large set of such clauses, each of which casts a "vote" for a particular class. These votes are then aggregated, and the overall prediction is determined by the balance of evidence provided by all the clauses together. This ensemble-like mechanism allows the model to capture subtle, complex patterns while still maintaining a form that remains human-interpretable.

We used the MultiClassTsetlinMachine from pyTsetlinMachine Python module and utilised the in-built *bit-per-feature binarization* to binarize the data [11]. This method discretizes continuous variables into a fixed number of bins, encoding each bin as a separate binary feature. This transformation ensures compatibility with TM's binary input format. The original bin values are stored separately to correctly identify the real values of the features corresponding to the clauses.

After binarization, an 80-20 train-test split was applied and the model was trained with appropriate hyper-parameters (i.e., the number of clauses, threshold, and specificity).

Our target variable represents stroke subtypes (a binary classification task) and the TM generated 50 clauses for each class. To identify the most influential clauses per class, we analysed their voting weights, which reflect how frequently a clause contributes to a particular class prediction. We selected the top clauses based on these weights to further enhance interpretability and explainability and to reduce redundancy, with two filters:

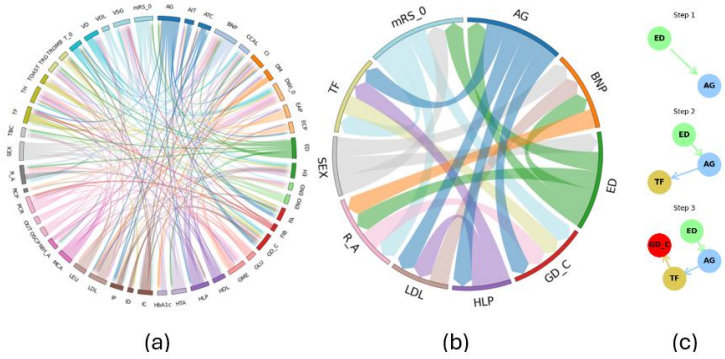

Figure 1. Causal graphs: (a) with all the features, (b) top 10 features using composite score of degree and betweenness centralities and (c) deconstructed specfic causal path

- *Bias Check*: We excluded clauses that were overwhelmingly positive or negative for a single class to avoid skewed interpretations.
- *Redundancy Check:* Clauses that appeared identically in both classes of the outputs were removed, as they introduce ambiguity in the interpretation of feature impact.

After filtering, we retained the distinct and unbiased clauses for each class with the highest voting weights. These clauses form the basis for interpreting how specific combinations of biomarker presence or absence influence the classification of stroke subtypes.

III. RESULTS AND DISCUSSION

Based on the process explained in the methodology section, our final goal is to obtain the top clauses for each of the output classes. To simplify further, we retrieve the most important features for each class as well as the information whether their absence or presence is important for either class.

In this section, we discuss the results of both the causal AI and the Tsetlin Machine.

A. Results of Causal inference

We extract the list of top nodes/features using the *composite centrality*, as defined in the methodology section. The causal Directed Acyclic Graph (DAG) connections comparing original features and the extracted top 10 features using causal inferences are shown in Figure 1.

The first graph (Figure 1a) presents the complete set of features and biomarkers included in the dataset. Because all variables and their interconnections are displayed at once, the result is a complex and visually dense network that makes it difficult to distinguish which biomarkers play the most critical roles. In contrast, the second graph (Figure 1b) focuses only on the top 10 most influential features, as identified through our causal inference procedure using the composite centrality score. This reduced network provides a much clearer picture of the variables that exert the strongest influence on stroke outcomes, allowing clinicians and researchers to focus on the most relevant biomarkers. To further illustrate how causal inference can assign importance to a feature, even when the connection to the target is indirect, the right-hand panel (Figure 1c) zooms in on a specific causal path. In this example, the feature *age (ED) in Figure 1b* does not connect directly to the target variable, GD-C, which represents the type of stroke. Instead, its influence is mediated through an intermediate biomarker, *AG* (prior use of antiplatelet drugs), which then affects *TF* (treatment to dissolve blood clots), and only at that point does the causal chain reach GD-C. This breakdown demonstrates how a variable can still be considered highly important when it contributes to the target outcome through a series of intermediate links, rather than through a direct relationship as well as to trace and understand how each node in the causal graph contributes to the target outcome, whether through direct or indirect pathways.

The top features/biomarkers identified by the causal model and their significance in the context of stroke related literature is summarized in Table 1 below.

TABLE I. MOST IMPORTANT BIOMARKERS AS PER CAUSAL MODEL

Feature	Description	Significance
BNP	Blood test to help diagnose heart failure	A strong indicator for cardiac stress, important for stroke diagnosis/prognosis
AG	Prior use of antiplatelet drugs	Aligns with existing clinical evidence that such medications reduce the risk of recurrent stroke
ED	Age of the patient	A critical determinant of stroke severity and recovery potential
HLP	Abnormally high levels of lipids (fats)	Associated with increased stroke risk; important for stroke prevention strategies
LDL	Bad cholesterol	Linked to atherosclerosis and subsequent cerebrovascular events; a key modifiable risk factor
R_A	Degree of disability after a stroke at discharge	Reflects the immediate functional outcome post-stroke; serves as a proxy for the effectiveness of acute care

mRS_0	Baseline disability in daily activities	Predictive of post-stroke recovery trajectories
SEX	Gender of the patient	Reflects gender effect in stroke prognosis and prevention
TF	Treatment to dissolve blood clots	Highlights a critical role of emergency treatments in improving stroke outcomes
GD_C	Category of the stroke type (target)	Classification of stroke types; target of this study

As can be seen from the *significance* column in Table 1, the causal model validates known clinical associations. Additionally, it also captures nuanced interdependencies among biomarkers by providing the strength of connections between them (i.e., node connection strengths calculated using *composite score* as described in the methodology section).

The model's ability to prioritize features with both statistical and clinical relevance strongly supports its potential application in decision support systems for stroke management.

B. Results of Tsetlin Machine

As previously mentioned, a TM produces human-readable clauses (e.g., *if A and not B, then class X*). After applying the model to the top features identified through causal inference, we derive such clauses for our target variable, the type of stroke.

Figure 2 provides a visual depiction of the clauses. In this illustration, pink cells indicate the absence of a feature for the corresponding class shown at the bottom, while light green cells represent its presence. Each feature's value range is displayed within its respective cell. The feature SEX is binarized, with 0 → female and 1 → male.

The clause for Ischemic stroke would then be:

> *If the modified Rankin Scale (mRS_0) score is greater than 2.67,* ***and*** *the LDL level is between 71 and 117 mg/dL,* ***and*** *the patient's age is not greater than 56 years,* ***and*** *the BNP level is not between 550 and 1123 pg/mL,* ***then*** *the predicted outcome is Ischemic stroke.*

Which in logic notation is:

$$IF\ (mRS_0 > 2.67) \wedge (71 < LDL > 117) \wedge (Age < 56) \wedge \neg(550 < BNP < 1123) \rightarrow ISCHEMIC$$

Such human-readable clauses, with well-defined value ranges for each feature or biomarker influencing the output classes, could become particularly valuable in clinical settings.

In terms of clinical research, they enhance model transparency, enabling researchers to validate findings against existing biomedical knowledge and uncover novel associations. This interpretability can help bridge the gap between data-driven models and domain expertise. Furthermore, such clauses can inform the design of prospective studies and contribute to the development of explainable clinical decision support tools.

Finally, having transparency in clinical decision-making would benefit effective patient communication, helping individuals understand prevention strategies and treatment options.

IV. Conclusion and Future Work

The findings presented here are preliminary and require further refinement. A key priority is to acquire additional

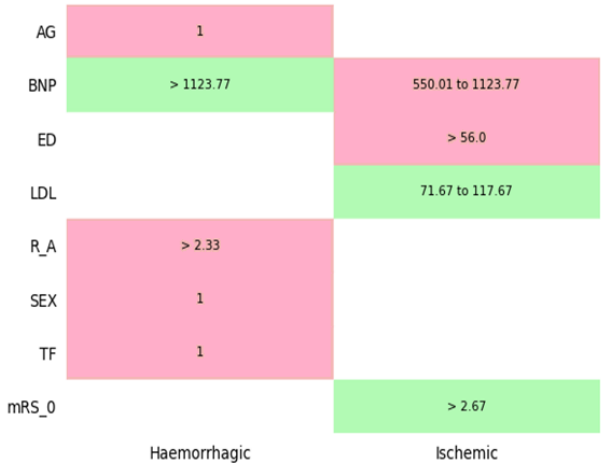

Figure 2. Visual representation of Tsetlin Machine clauses identified for the target with most important biomarkers.

data and repeat the analysis to ensure the robustness of the results. We are in the process of obtaining a more comprehensive dataset, which will include recent records of stroke patients.

To further strengthen the robustness of the results, the next steps are broadly categorized into two areas: one focusing on Causal AI and the other on rule extraction using the Tsetlin Machine.

A. Causal AI

To ensure the accuracy of the causal graphs, it is essential to correctly capture the directionality of the relationships. Achieving this will require deeper domain expertise and a thorough analysis of how various biomarkers interact.

Additionally, it is vital to conduct *what-if* scenario simulations based on the discovered causal relationships within the feature space. These *in-silico* experiments will

enable us to explore how changes in feature values, whether hypothetical or novel, might influence stroke prognosis, without the need for new empirical data.

B. Tsetlin machine

While our current model achieves an overall accuracy of approximately 80%, a closer examination of its performance metrics reveals a notable imbalance. Specifically, the F1-score for Class 0 (the majority class) reaches 0.88, whereas the F1-score for Class 1 (the minority class) drops sharply to just 0.15. This large disparity highlights that, although the model performs well in predicting the dominant class, it struggles to correctly identify cases that belong to the less frequent class. In practice, this means that the model fails to capture a substantial proportion of minority class instances, which may correspond to clinically critical or rare conditions. The root cause of this problem is the class imbalance present in the dataset, where examples of one stroke subtype greatly outnumber the other. We anticipate that the inclusion of additional patient records in our forthcoming dataset will help mitigate this imbalance by providing a more even distribution of classes.

It is also important to emphasize that a Tsetlin Machine (TM) differs fundamentally from many classical machine learning models. Instead of optimizing a global error function, the TM relies on a frequency-driven clause learning mechanism in which the prevalence of certain patterns directly affects the clauses it learns. While this makes the model efficient and interpretable, it also means that it tends to favor patterns associated with the majority class, often at the expense of learning sufficient rules for the minority class. This characteristic can amplify the effects of class imbalance, as seen in our results.

Nevertheless, in the context of biomedical datasets (where imbalanced class distributions are common) this bias does not necessarily negate the model's clinical utility. Optimizing for the majority class can still yield valuable insights, as the most prevalent stroke subtype remains a major focus of clinical diagnosis and treatment. However, achieving reliable detection of minority cases is equally critical, as these often represent the most challenging and high-risk scenarios. Addressing this imbalance in future work will therefore be essential, ensuring that the TM captures meaningful patterns for both majority and minority classes without sacrificing interpretability.

These facts also do not diminish the importance of accurately identifying minority class instances, which often represent critical or rare conditions. To address this, we are actively exploring various strategies (e.g., resampling, decision threshold tuning, etc.) to improve the model's ability to generalize and perform equitably across both classes. These efforts are guided by domain expertise to ensure that learned patterns are meaningful and to prevent the model from learning artifacts of the data rather than true signals.

Additionally, binarization must be approached with greater care. It is important to ensure that the binning of biomarkers identified as significant by the Tsetlin Machine aligns with domain knowledge and statistical distribution. For example, consider serum Vitamin D levels, which typically range from 0 to 100 ng/mL. Clinical guidelines define severe deficiency as levels below 10 ng/mL, deficiency as below 20 ng/mL, insufficiency between 20–30 ng/mL, and sufficiency as levels above 30 ng/mL. If all values below 30 ng/mL were grouped into a single bin (e.g., bin 0), this would obscure critical clinical distinctions between mild insufficiency and severe deficiency. Such coarse binning could reduce the model's ability to detect meaningful health risks associated with different deficiency levels.

ACKNOWLEDGMENT

The authors would like to thank the Instituto de Salud Carlos III_ICIII RICORS-ICTUS network (grant number RD24/0009/0017) and Xunta de Galicia (grant number: IN607A2022/02) for providing the resources to carry out this work.

REFERENCES

[1] Ministry of Health, "Annual Report of the National Health System 2023," Government of Spain, Aug. 2024. [Online]. Available: https://www.sanidad.gob.es [retrieved: Sep., 2025].

[2] Eurostat, "Causes of death statistics," *Statistics Explained*, European Commission, Mar. 2025. [Online]. Available: https://ec.europa.eu/eurostat/statistics-explained [retrieved: Jun., 2025].

[3] B. Mirza, W. Wang, J. Wang, H. Choi, N. C. Chung, and P. Ping, "Machine learning and integrative analysis of biomedical big data," *Genes*, vol. 10, no. 2, pp. 87, 2019. [Online]. Available: https://doi.org/10.3390/genes10020087 [retrieved: Aug., 2025].

[4] P. Sanchez, J. P. Voisey, T. Xia, H. I. Watson, A. Q. O'Neil, and S. A. Tsaftaris, "Causal machine learning for healthcare and precision medicine," *Royal Society Open Science*, vol. 9, no. 7, pp. 220638, 2022. [Online]. Available: https://doi.org/10.1098/rsos.220638 [retrieved: Jul., 2025].

[5] O. C. Granmo, "The Tsetlin Machine – A game theoretic bandit driven approach to optimal pattern recognition with propositional logic," *arXiv preprint arXiv:1804.01508*, 2018. [Online]. Available: https://arxiv.org/abs/1804.01508 [retrieved: Sep., 2025].

[6] G. T. Berge, O. C. Granmo, T. O. Tveit, B. E. Munkvold, A. L. Ruthjersen, and J. Sharma, "Machine learning-driven clinical decision support system for

concept-based searching: A field trial in a Norwegian hospital," *BMC Medical Informatics and Decision Making*, vol. 23, no. 5, pp. 1–12, 2023. [Online]. Available: https://doi.org/10.1186/s12911-023-02101-x [retrieved: Jun., 2025].

[7] M. Kalisch and P. Bühlmann, "Estimating high-dimensional directed acyclic graphs with the PC-algorithm," *Journal of Machine Learning Research*, no. 8, pp. 613–636, 2007. [Online]. Available: https://jmlr.csail.mit.edu/papers/v8/kalisch07a.html [retrieved: Aug., 2025].

[8] A. Wheeldon, A. Yakovlev, and R. Shafik, "Self-timed reinforcement learning using Tsetlin Machine," in *Proceedings of the 27th IEEE International Symposium on Asynchronous Circuits and Systems (ASYNC 2021)*, IEEE, 2021. [Online]. Available: https://arxiv.org/abs/2109.00846 [retrieved: Jun., 2025].

[9] S. Glimsdal and O.-C. Granmo, "Coalesced multi-output Tsetlin Machines with clause sharing," *arXiv preprint arXiv:2108.07594*, 2021. [Online]. Available: https://arxiv.org/abs/2108.07594 [retrieved: Sep., 2025].

[10] K. D. Abeyrathna, O.-C. Granmo, and M. Goodwin, "Adaptive sparse representation of continuous input for Tsetlin Machines based on stochastic searching on the line," *Electronics*, vol. 10, no. 17, pp. 2107, Aug. 2021. [Online]. Available: http://dx.doi.org/10.3390/electronics10172107 [retrieved: Jul., 2025].

[11] O. C. Granmo, et al., "pyTsetlinMachine [Computer software]," GitHub, n.d. [Online]. Available: https://github.com/cair/pyTsetlinMachine [retrieved: Jun., 2025].

Explainability Analysis for Skill Execution

Khatina Sari, Paul G. Plöger, and Alex Mitrevski
Department of Computer Science
Hochschule Bonn-Rhein-Sieg
Sankt Augustin, Germany
e-mail: khatina.sari@smail.inf.h-brs.de, {paul.ploeger; aleksandar.mitrevski}@h-brs.de

***Abstract*—Explainability holds significant importance for autonomous robots deployed in human-centered situations, particularly when errors occur during execution. In the context of robot action, it is important to consider various levels and types of explainability. The social dimension of Artificial Intelligence (AI) and robotic explanations, which highlights how they affect social interaction, values, and decision-making, has received little to no attention in prior research. With a particular emphasis on item handover, we hypothesize that users prefer systems with explanations and that explanations in natural language are more appealing than heatmaps. A user study, involving participants from diverse backgrounds and levels of expertise, is conducted to evaluate different levels and preferred types of explainability. The study results support our hypotheses and offer additional valuable information for future system development.**

Keywords-Explainable Artificial Intelligence; Natural Language Processing; Heatmaps; Human-Robot Interaction.

I. INTRODUCTION

There have been notable developments in the disciplines of Artificial Intelligence (AI) and robotics in recent decades, which are both largely affiliated. Future robotics systems are anticipated to be far more advanced and adaptable as AI and robotics continue to grow. Rule-based systems, also referred to as white-box artificial intelligence, place an emphasis on transparency, making their logic processes clear and accessible to users. On the other hand, black-box AI, such as neural networks, often does not specify its decision-making process. Therefore, researchers are actively refining the interpretability of black-box AI, which can be used to improve transparency in robot actions, especially when failures occur [1-5].

Some challenges in Human-Robot Interaction (HRI) necessitate transparent communication. Varying user knowledge and expectations pose challenges in maintaining the right level of detail in the explanations. Another challenge is to determine the most effective explanation format for each user [6][7]. Explainability can be classified as local (usually focused on a single input dataset), global (describing how a model behaves generally), model-specific [8] (limited to particular model classes), model-agnostic [9] (may be local or global and independent of machine learning models), and counterfactual [10] (offering an alternate input scenario that would have produced a different model prediction). Meanwhile, there are three common levels of explainability [8]: low-level (which includes techniques like linear model coefficients or feature importance scores), medium-level (which delves deeper into how specific features impact the model's predictions), and high-level (which highlights intricate decision-making processes within the model).

This paper is focusing on robot object handover tasks, with the intention to enhance user understanding and trust in robot actions. A user study was conducted to evaluate the effectiveness of multiple levels of explainability in such tasks. This study aims to encourage innovation in autonomous robotics by providing access to more adaptable, flexible, and user-centered systems.

The remainder of this paper is organized as follows. Section II offers an overview of literature related to the challenging topic this study addresses. Section III describes the general approaches used in our methodology. Section IV outlines our experimental results, both qualitative and quantitative, as well as hypothesis testing. Section V summarizes our findings and includes possible future work.

II. RELATED WORK

Transparent or white-box models refer to algorithms that provide users with both the end decision and a summary of the steps used to get there. One of the most common methods used for this is Bayesian network [11][12]. However, this method often requires substantial manual effort from users to explore the robot's behavior [13]. It lacks scalability and generalizability because it involves hand-annotating every domain-specific context up front, which hinders application to new circumstances.

On the other hand, opaque or black-box models are machine learning models that are difficult to explain and understand by experts in practical domains [14][15]. These models include random forest, support vector machine, multilayer neural network, etc. One of the ways to obtain information from such models are to use post-hoc interpretability. Although this approach provides useful information for end users, it often does not clarify precisely how a model works. Therefore, a more thorough analysis of a better strategy for building trust, reliance, and performance for human-AI teams needs to be conducted.

The need for user-centered design practices when creating explanations for AI systems was emphasized by [16]. They suggest involving users in the AI system design process

 ISBN: 978-1-68558-318-7

through user studies, interviews, and feedback sessions to understand their needs, mental models, and expectations. Even so, they primarily focused on design practices and guidelines for creating user experiences in explainable AI systems and did not delve deeply into technical solutions or algorithms to achieve explainability. As a result, the technical aspects of implementing the proposed guidelines may require further exploration.

In Human-Robot Collaboration (HRC), human workers should have the ability to naturally converse with robots, since they are the most crucial members of any HRC team. According to [17], while there are currently few means of communication between human workers and robots, gesture recognition has long been used as an efficient human-computer interaction. In conclusion, they believe that HRC will operate in a safer environment if a depth sensor and body-model technique are combined to track human movements.

As part of the machine learning adaptation in the robot's motion planning, our approach proposes the utilization of a neural network. This is an alternative approach to the genetic algorithm utilized by [14]. The adjustment in methodology highlights our dedication to investigating different and practical approaches that may result in improved responsiveness and flexibility of robotic systems in dynamic settings. In addition, inspired by [16] user-centric principles, we conducted a user study to uncover user preferences regarding different approaches in robot motion planning. Our questionnaire aims to uncover user preferences regarding the different approaches employed in robot motion planning, shedding light on which method resonates more effectively with particular users.

III. APPROACH

The scope of our study concentrates on the usage of autonomous robots for object handover tasks from robot to human, an important use case that requires an effective explanation strategy. Giving our Toyota Human Support Robot (HSR) a skill set that corresponds to different levels of explainability—or, in some cases, no explainability at all—is the current challenge at hand. Our explainability analysis for skill execution takes into account a number of important factors, one of which is the recognition that explainability in our case is inherently local.

A. Proposed Approach

The current approach used in our robot to determine the handover position is done by factoring in context-dependent (based on the posture of the detected person) and context-independent (static; based on the context-dependent outcome). However, the handover position in a context-independent approach does not consider any surrounding environment variables; thus, we propose to train a neural network to dynamically set the end-effector position based on the values obtained from the 3D bounding box. By allowing the neural network to generate random handover positions, we can collect input-output pairs dataset that can be used to fine-tune the model until it can automatically generate optimal handover positions based on the user's needs. This strategy would increase the effectiveness and usability of the robotic system. Regrettably, a prolonged mechanical issue in our Toyota HSR has forced us to delay the implementation of our neural network interpretation. Upon its resumption of operations, we shall resume our work and implement our planned approach.

B. Explainability Setup

One of the primary concerns that drives our research is how to determine the robot's reasoning behind certain decisions, especially why it stops at a specific point in relation to the detected human position during object handover. To carry out this research, an advanced built-in program created by [18] is used, which generates a 3D bounding box to locate the detected person in front of the robot. It follows the right-handed coordinate system, which includes the depth (x-axis), horizontal (y-axis), and vertical (z-axis). Once the person is detected, their position will be determined; in our case, there are three possible positions: standing, sitting, and lying down.

Within our research framework, several notations play an important role in influencing how we perceive the spatial connection between humans and the robot during the handover task. Figure 1 illustrates the configuration in which W_p represents the robot's end-effector location where the object is held, W is the robot's base frame, B denotes the bounding box, and p is the relative position between the end effector and the center point of the bounding box.

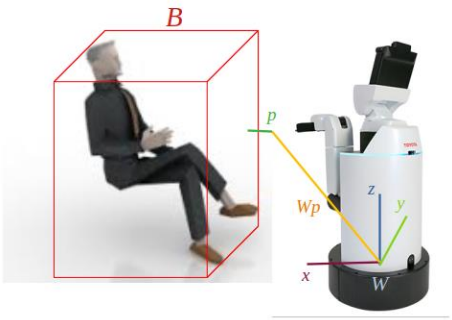

Figure 1. Illustration of the parameters on handover skill.

Logical predicates describing the requirements for a successful handover interaction are adopted from [19] to define the success preconditions. The predicates include $in\_front\_of_{x,y}(p,B)$, $far\_in\_front\_of_{x,y}(p,B)$, $behind_{x,y}(p,B)$, $far\_behind_{x,y}(p,B)$, $above_{x,y}(p,B)$, $below_{x,y}(p,B)$, and $centered_{x,y}(p,B)$. Using the success preconditions, the natural language explanation for each position is generated manually, as shown in Tables I-III.

In addition to manual natural language translation, ChatGPT 3.5 [20] is employed to generate automated translation and evaluate the results using the Bilingual Evaluation Understudy (BLEU) score [21]. The first few initial tests did not produce close translations to the manual translation. Therefore, more detailed definitions of each logical expression were provided, as well as separating each predicate that consists of two or more coordinates; for

example, $centered_{x,y}(p,B)$ becomes $centered_x(p,B) \wedge centered_x(p,B)$. The outcome of the last iteration was then used for assessment.

TABLE I. PRECONDITIONS FOR STANDING POSITION

Types	Success Preconditions
Logical Predicates	$centered_{y,z}(p,B) \wedge in\_front\_of_x(p,B) \wedge \neg centered_x(p,B) \wedge \neg below_{x,y}(p,B) \wedge \neg behind_{x,y}(p,B) \wedge \neg far\_behind_{x,y}(p,B) \wedge \neg above_{x,y}(p,B) \wedge \neg in\_front\_of_y(p,B) \wedge \neg far\_in\_front\_of_y(p,B)$
Natural Language	The robot's arm should be in front of and centered around a person (corresponding to the person's height and width). It should not be behind, above, beneath, or to the right/left of a human.

TABLE II. PRECONDITIONS FOR SITTING POSITION

Types	Success Preconditions
Logical Predicates	$centered_{y,z}(p,B) \wedge in\_front\_of_x(p,B) \wedge \neg centered_x(p,B) \wedge \neg below_{x,y}(p,B) \wedge \neg behind_{x,y}(p,B) \wedge \neg far\_behind_{x,y}(p,B) \wedge \neg above_{x,y}(p,B) \wedge \neg in\_front\_of_y(p,B) \wedge \neg far\_in\_front\_of_{x,y}(p,B)$
Natural Language	The robot's arm is positioned in front of and around the middle of a sitting person (according to the person's height and width). It is not behind, above, beneath, and to the right or left of the person.

TABLE III. PRECONDITIONS FOR LYING DOWN POSITION

Types	Success Preconditions
Logical Predicates	$above_{x,y}(p,B) \wedge centered_y(p,B) \wedge \neg centered_z(p,B) \wedge \neg below_{x,y}(p,B) \wedge \neg behind_{x,y}(p,B) \wedge \neg far\_behind_{x,y}(p,B) \wedge \neg in\_front\_of_y(p,B) \wedge \neg far\_in\_front\_of_{x,y}(p,B)$
Natural Language	The robot's arm is positioned above and centered around the person's width. It is not below or around their head or feet. It should not extend all the way to the opposite side from where a robot is standing next to.

BLEU provides a quantitative measure by comparing the output of machine translation systems (candidate translation) against reference translations, offering insights into the degree of overlap in n-gram or word sequences with human-generated counterparts [21]. The length of candidate sentences that are shorter than the reference phrases is penalized in the BLEU metric (Brevity Penalty), which is based on the modified n-gram precision measure. The following formula determines the BLEU score:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} \frac{1}{N} \cdot logP_n\right), \qquad (1)$$

where BP = Brevity Penalty and P_n = Precision for n-gram.

The Natural Language Toolkit (NLTK) [22] and spaCy [23] are used in our BLEU score computation to provide an unbiased evaluation of machine-generated translations. The translation produced by ChatGPT 3.5 (as a candidate translation) is compared with our original translation (as a reference). The results of the BLEU score for each translation performed by ChatGPT in comparison to the manual translation are presented in Table IV.

TABLE IV. BLEU SCORE OF CHATGPT 3.5 TRANSLATION

No.	Position	BLEU Score
1	Standing	0.85
2	Sitting	0.81
3	Lying Down	0.88

The final translation output from ChatGPT 3.5 provides a good starting point for future developments. Despite the fact that the translations produced by the first few iterations were not satisfactory, adding further specific information made it generate a translation that was similar to the one that was done manually. The key realization is that it is possible to train models, like ChatGPT, to translate technical terminology into natural languages effectively.

When it comes to interpreting the neural network's decisions about handover position, Grad-weighted Class Activation Mapping (Grad-CAM) [24] integration shows itself to be an effective tool for insight. It offers a transparent and insightful lens into the decision-making processes of complex models. Grad-CAM fills this gap by giving an illustration of the areas in the input data that have a major impact on a certain outcome. Unfortunately, the problem with our Toyota HSR prevented us from implementing this method. Despite this obstacle, a previously collected dataset from our research team [25] was leveraged, and the video content was edited to achieve the same heatmap effect (as seen in Figure 2). This decision allowed us to simulate and observe the intended outcomes, ensuring the continuity of the research despite the technical constraints.

Figure 2. Additional heatmaps on one of the handover scenarios.

The dataset, which includes relevant information but lacks explanations, was then extended by adding explanations in both heatmap and natural language formats. This improvised solution allows us to proceed with our user study within the designated timeframe, preserve the research objectives, and ensure the timely execution of the study.

C. Experimental Design

In our comprehensive user study aimed at investigating user preferences in interacting with AI-based or robotic systems, two distinct hypotheses were formulated to guide our research. The first hypothesis is that users have a preference for systems that offer explanations while they are using them. The second hypothesis is about the preferred explanation format among users; in particular, we hypothesize that people

prefer explanations in natural language over alternative visualization techniques like heatmaps.

Our user study adopts a mixed-methods strategy to gather quantitative data and qualitative insights through surveys in order to experimentally validate our hypotheses. After being presented with simulated robotic interfaces that include heatmaps and natural language explanations, participants' preferences, satisfaction, and understanding were carefully examined.

Through selectively crafted survey questions, user experiences, preferences, and challenges are explored, allowing us to obtain insights into the factors that contribute to a positive or negative interaction. Additionally, scenarios that are meant to replicate real-world interactions were chosen by giving users experiences that were contextually appropriate and reflected the difficulties and complexities of real-world circumstances. Ten videos and three different explanation varieties were presented to help construct a more comprehensive understanding of user preferences: no explanation, partial explanation using heatmaps, and detailed explanation using natural language. In order to prevent any potential biases, 8 out of 10 videos were purposefully presented in a random order. Following every video, participants were asked to rate how confident they were in their understanding of the robot decision-making process.

IV. EXPERIMENTAL RESULTS

Our user study involved a total of 33 participants, ages ranging from 18 to 40 years old, education ranging from high school to Ph.D., and different academic and professional backgrounds. Our participants' demographic profiles show a variety of age groups, gender identities, levels of education, and fields of study. This diversity attempts to determine whether there is any relationship between the preferred explanation technique and the educational background.

A. Quantitative Analysis

In terms of the participants' experiences and expectations in the realms of robotic systems and Artificial Intelligence (AI), 75.8% of them have prior hands-on experience with robotic systems, while an overwhelming 84.8% are familiar with AI or machine learning in their practical lives. In a survey on comfort levels, 72.7% of the respondents said they felt uneasy when AI systems made decisions without providing an explanation, highlighting the significance of transparency.

In our scenario-based questions, two identical videos served as starting points. The first was without explanation, whereas the second included a natural language explanation. The majority indicated that they were unclear about the robot's action in the first video, though it was a successful object handover scenario. However, the participant's confidence level improved after watching the second video, which revealed a positive beginning. Table V summarizes participants' confidence levels after eight more videos were shown in a random order. It reveals that individuals feel more confident when they are given an explanation of how the robot makes decisions. Less than 40% of the participants felt confident about their understanding of the robot decision-making process in the three videos without an explanation, in both successful and unsuccessful handover scenarios. More than 50% of the participants in the two videos where heatmaps were used as an explanation type expressed confidence in the successful handover scenario. However, in the case of an unsuccessful handover, only 34.6% of participants reported feeling confident. With natural language explanations, on the other hand, 48.4% of those surveyed expressed confidence in the unsuccessful scenarios. In the successful scenario, over 80% of the participants expressed confidence and none of them indicated lack of confidence.

TABLE V. AN OVERVIEW OF PARTICIPANTS' CONFIDENCE LEVEL

Video	Outcome	Explanation Type	Confidence Level (%)				
			5	4	3	2	1
3	Succeed	None	9.1	24.2	48.5	18.2	0.0
4	Succeed	Heatmap	24.2	27.3	36.4	12.1	0.0
5	Failed	None	12.1	15.2	30.3	24.2	18.2
6	Failed	Natural Language	24.2	24.2	15.2	15.2	21.2
7	Succeed	None	3.0	18.2	15.2	36.4	27.3
8	Succeed	Natural Language	27.3	57.6	15.2	0.0	0.0
9	Failed	Natural Language	24.2	24.2	18.2	27.3	6.1
10	Failed	Heatmap	18.2	21.2	30.3	27.3	3.0

To conclude, compared to visual explanation (using a heatmap), natural language explanation improves their confidence by over 30% (shown in Figure 3).

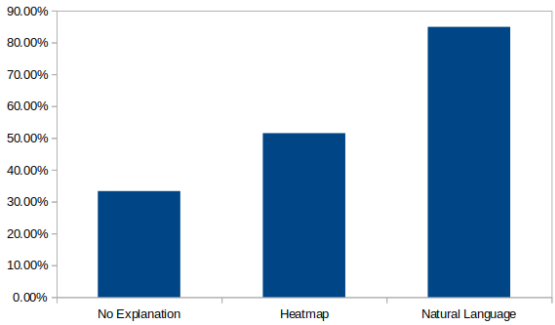

Figure 3. Participants' overall confidence in understanding the robot decision-making process.

B. Hypotheses Testing

We conducted the first hypothesis test to investigate users' preferences regarding the type of videos when seeking information. The hypothesis aimed to determine whether users prefer videos with explanations over videos without explanations. The participants were presented with the question "Which type of video do you prefer when seeking information?" and the response options: videos with explanation, without explanation, and depending on the context.

A chi-square test [26] for independence is employed to analyze the association between the type of video and user preference, where H_0 = no preference difference and H_1 = there is a preference for videos with an explanation. If the p-

value of a given dataset is less than 5%, the null hypothesis is rejected because it is assumed that there is a preference difference among the options. To calculate the p-value using chi-square formula (2), the observed value (O) needs to be identified first, which represents the actual counts derived from the sample, and the expected value (E), which represents the values of each category in the event that there was no preference difference between all categories. E is obtained by dividing the total number of observed values by the number of categories. The following calculation can then be used to get its chi-square statistic (χ^2) based on the observed and expected values:

$$X^2 = \sum \frac{(O-E)^2}{E}. \quad (2)$$

The result, along with the degrees of freedom (df), which is a number representing how much variation is involved in the research (n) minus 1,

$$df = n - 1, \quad (3)$$

is used to calculate the p-value from the chi table.

Our observed and expected values based on the survey results are displayed in Table VI. The total observed values—33 in this case—and the number of categories—3 in this case—are then used to compute the expected values, yielding the value E = 11.

TABLE VI. THE OBSERVED AND EXPECTED VALUES

User Preference	O	E	$O-E$	$(O-E)^2$
With Explanation	22	11	11	121
Without Explanation	2	11	-9	81
Depend on the Context	9	11	-2	4

These observed and expected values were used to calculate the chi-square statistic, which was then used to test the hypothesis. The result yielded χ^2= 28.1; with df = 2, the resulting p-value was 0.0000008. Since the p-value is less than α = 5% or 0.05, it is determined that the null hypothesis is rejected.

The second hypothesis is tested based on two identical videos with two distinct explanations—one using a heatmap (video 4) and the other using natural language (video 8). Participants were asked to choose which of the two videos gave them a better understanding of the robot decision-making process. Participants who selected video 8 are considered to prefer the natural language explanation. A one-sample proportion test (Z) [27] is employed to analyze whether the proportion of users who prefer video 8 differs significantly from 50% (no preference). The null hypothesis (H_0) assumed no preference difference, while the alternative hypothesis (H_1) assumed a preference for videos with natural language explanation.

To conduct the test, we need to estimate the proportion $\hat{p}$ as:

$$\hat{p} = \frac{x}{n}, \quad (4)$$

where x is the number of participants who have chosen video 8 and n is the total number of participants. After that, the test statistic can be calculated with the following formula:

$$Z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \quad (5)$$

where p_0 is the pre-specified value; in this case, it is 50% to indicate that if half of the total participants chose video 8, there is no significant preference for that particular video. From there, the calculated Z-value is compared with critical values, which can be obtained from the Z table, from the standard normal distribution. Given that the sampling distribution of our data is a normal distribution with a significant value of 0.05, the critical values are in a range of -1.96 to 1.96. Based on the result of our survey, a one-sample proportion test was calculated with x = 23 and n = 33, which yielded a Z-value of 2.46. Because the Z-value is larger than the maximum critical value, the null hypothesis is rejected.

A post hoc sensitivity analysis [28] was conducted to evaluate the statistical power of our study. Cohen's w,

$$w = \sqrt{\sum \frac{(p_i-p_{0i})^2}{p_{0i}}}, \quad (6)$$

where p_i is the observed value in category i and p_{0i} is the expected value under the null hypothesis in category i, is used to measure the effect size for the chi-square test of the first hypothesis. The thresholds are 0.10 for a small effect, 0.30 for a medium effect, and 0.50 for a large effect. The result yielded w = 0.75, which represents a large effect.

Furthermore, we assess the effect size for the one-sample proportion test of the second hypothesis with Cohen's h,

$$h = 2\,x\left(\arcsin\left(\sqrt{p_1}\right) - \arcsin\left(\sqrt{p_2}\right)\right), \quad (7)$$

where p_1 and p_2 are the two proportions being compared. The thresholds are 0.20 for a small effect, 0.50 for a medium effect, and 0.80 for a large effect. From our user study result, 23 out of 30 participants preferred video with natural language explanation; thus, p_1 = 69.7%. Then we compare it with p_2 = 50% for the proportion that shows no preference difference. The result yielded h = 0.40, which indicates a moderate effect size.

C. Qualitative Analysis

As proven in our hypothesis 2, natural language explanations are preferable to heatmaps. In order to evaluate it on a qualitative level, the participants were asked why they

preferred one type of explanation over the other, and the majority of them responded that they preferred natural language because it is easier to understand and more elaborate. In addition, they believe that natural language explanations can be enhanced by an audio or speech component.

They were then asked to imagine a situation in which they would favor a different kind of explanation than the one they had previously selected. Those who have chosen natural language say that they prefer heatmaps when a robot performs a simple task, interacts with static objects, or is in a simulation. On the other hand, those who have chosen heatmaps say that they prefer natural language when failure occurs, when the robot is in a dynamic environment, or when the user has no background knowledge about the system.

When asked to imagine a situation in which they would prefer to have no explanation at all, the majority of respondents believe that in a straightforward or routine task that is repeated, there is no need for an explanation because the rationale is obvious. While some claim that they cannot think of any situation in which it is preferable not to have an explanation, others highlight this point by stating that, even in tasks that appear straightforward, having an explanation is desirable since it provides a clear reasoning behind the robot's chosen action.

V. CONCLUSION AND FUTURE WORK

Our user study results supported our hypotheses, offering statistical evidence that users do, in fact, prefer explanations when interacting with robotic systems. These findings highlight that providing explanations improves users' trust and understanding of robot systems. Although the study demonstrates a clear preference for explanations in natural language as opposed to heatmap visualizations, respondents express a preference for heatmaps or no explanations at all when the robot is performing regular or routine tasks. This tendency implies that, in situations they are familiar with, participants think that the visual representations of the heatmaps are sufficient or that perhaps they prefer them more when the tasks are simple and require no extra information. Due to the wide range of participant preferences, flexible communication strategies that take into account varying user expectations and levels of experience with certain robotic tasks are necessary.

Even though the results suggest that users prefer systems that provide explanations over those that do not, it is important to acknowledge a potential bias in how this hypothesis was tested. The question itself highlights the presence or absence of an explanation, which might have led participants to gravitate toward the condition with explanations, independent of their actual utility in decision making. Future studies should aim to mitigate this bias by embedding explanations in more naturalistic tasks where the usefulness of the explanation emerges organically rather than being made explicit to participants.

While our findings indicate that participants preferred natural language explanations, it is important to recognize that this result may partly reflect differences in interpretability between formats. Natural language requires little effort to process, whereas heatmaps demand additional interpretation and prior familiarity. This asymmetry may have disadvantaged the heatmap condition. To address this imbalance, future studies should explore providing training or familiarization with visual explanations, refining visualization design to reduce cognitive effort, or presenting hybrid formats that combine textual and visual elements for complementary strengths.

Further studies could explore automating the translation of scientific terms into natural language to provide explanations for nonexpert users. To implement audio explanations effectively, future work may explore the integration of speech synthesis technologies or Natural Language Processing (NLP) models specialized in generating spoken content. Additionally, exploring the potential of machine learning techniques, such as reinforcement learning, could contribute to optimizing explanation selection. This way, the system could learn over time which combination of explanation modalities yields the most positive user responses or facilitates optimal task performance.

ACKNOWLEDGMENT

The authors would like to express gratitude to all of the participants who willingly participated in the research, contributing their time and insights. Without their cooperation, this study would not have been possible.

REFERENCES

[1] O. Loyola-Gonzalez, "Black-box vs. white-box: understanding their advantages and weaknesses from a practical point of view," IEEE Access, vol. 7, pp. 154096–154113, 2019.

[2] W. Samek, T. Wiegand, and K.-R. M¨uller, "Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models," arXiv preprint arXiv:1708.08296, 2017.

[3] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," IEEE access, vol. 6, pp. 52 138–52 160, 2018.

[4] R. Guidotti, et al., "A survey of methods for explaining black box models," ACM computing surveys (CSUR), vol. 51, no. 5, pp. 1–42, 2018.

[5] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (xai) program," AI magazine, vol. 40, no. 2, pp. 44–58, 2019.

[6] A. Holzinger, "From machine learning to explainable ai," in 2018 world symposium on digital intelligence for systems and machines (DISA). IEEE, pp. 55–66, 2018.

[7] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," Journal of Artificial Intelligence Research, vol. 70, pp. 245–317, 2021.

[8] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," arXiv preprint arXiv:2006.11371, 2020.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," arXiv preprint arXiv:1606.05386, 2016.

[10] R. M. Byrne, "Counterfactuals in explainable artificial intelligence (xai): evidence from human reasoning," in IJCAI, pp. 6276–6282, 2019.

[11] M. Lomas, et al., “Explaining robot actions,” in Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot, pp. 187-188, 2012.

[12] D. Das, S. Banerjee, and S. Chernova, “Explainable ai for robot failures: generating explanations that improve user assistance in fault recovery,” in Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, pp. 351–360, 2021.

[13] O. Amir, F. Doshi-Velez, and D. Sarne, “Summarizing agent strategies,” Autonomous Agents and Multi-Agent Systems, vol. 33, pp. 628–644, 2019.

[14] M. Mucientes and J. Casillas, “Quick design of fuzzy controllers with good interpretability in mobile robotics,” IEEE Transactions on Fuzzy Systems, vol. 15, no. 4, pp. 636–651, 2007.

[15] A. Alvanpour, S. K. Das, C. K. Robinson, O. Nasraoui, and D. Popa, “Robot failure mode prediction with explainable machine learning,” in 2020 IEEE 16th International Conference on Automation Science and Engineering (CASE). IEEE, pp. 61–66, 2020.

[16] Q. V. Liao, D. Gruen, and S. Miller, “Questioning the ai: informing design practices for explainable ai user experiences,” in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–15, 2020.

[17] H. Liu and L. Wang, “Gesture recognition for human-robot collaboration: a review,” International Journal of Industrial Ergonomics, vol. 68, pp. 355–367, 2018.

[18] A. F. Abdelrahman, “Incorporating contextual knowledge into human-robot collaborative task execution,” Technical Report, H-BRS Sankt Augustin, 2020.

[19] A. Mitrevski, “Skill generalisation and experience acquisition for predicting and avoiding execution failures,” Ph.D. Dissertation, RWTH Aachen, 2023.

[20] OpenAI. (2023). ChatGPT (Oct 16 version) [Large language model]. [Online]. Available from:
https://chat.openai.com/chat. Retrieved: June 2024.

[21] M. Evtikhiev, E. Bogomolov, Y. Sokolov, and T. Bryksin, “Out of the bleu: how should we assess quality of the code generation models?” Journal of Systems and Software, vol. 203, p. 111741, 2023.

[22] S. Bird, “NLTK: the natural language toolkit,” in Proceedings of the COLING/ACL 2006 interactive presentation sessions, pp. 69-72. 2006.

[23] Y. Vasiliev, “Natural language processing with Python and spaCy: A practical introduction,” No Starch Press, 2020.

[24] R. R. Selvaraju, et al., “Grad-cam: visual explanations from deep networks via gradient-based localization,” in Proceedings of the IEEE international conference on computer vision, pp. 618–626, 2017.

[25] IROS 2022 HEART-MET Handover Failure Detection Challenge. Available from: https://codalab.lisn.upsaclay.fr/competitions/6757#learn_the_details-evaluation. Retrieved: September 2025.

[26] Pandis, N., “The chi-square test,” American Journal of Orthodontics and Dentofacial Orthopedics, vol. 150, no. 5, pp. 898–899, 2016.

[27] David I, Adubisi O, Farouk B, and Adehi M., “Assessing MSMEs growth through rosca involvement using paired t-test and one sample proportion test,” J Soc Econ Stat, vol. 9, no.2, pp. 30–42, 2020.

[28] Cohen, D. “Culture, social organization, and patterns of violence,” Journal of Personality and Social Psychology, vol. 75, no. 2, pp. 408–419, 1998, doi:10.1037/0022-3514.75.2.408.

RanXplain: Explaining Rankings in Recommendation Systems

Atakan Yilmaz, Nuray Eylul Erler, Emre Atilgan and Melisa Bal Aslan
Trendyol Group, Data Science
Istanbul, Turkey
Email: {atakan.yilmaz | eylul.erler | emre.atilgan | melisa.bal}@trendyol.com

***Abstract*—Recommendation systems are designed to rank items according to users' predicted interest. As these systems increasingly affect choices in domains like e-commerce and media, understanding the reasoning behind their rankings becomes essential. However, most existing approaches that explain recommendations focus on individual predictions, rather than explaining why one item is prioritized over another. To bridge this gap, this paper introduces RanXplain, an approach specifically designed to explain the ranking decisions produced by recommendation models. RanXplain operates as a separate machine learning model trained on pairs of items, using features that are derived from the original ranking model. The impact of different feature sets and model architectures on model performance is systematically investigated. Furthermore, a simulation based performance evaluation was presented on different breakdowns, specifically analyzing the proximity of item ranks and whether items belong to the same category to detect scenarios in which RanXplain yields superior performance. A practical insight is discussed regarding instances in which RanXplain fails to identify the ranking model's prioritization.**

Keywords-Recommendation System; Explainable AI (XAI); Machine Learning Explainability.

I. INTRODUCTION

Explainability in machine learning has become a cornerstone of responsible and trustworthy artificial intelligence, especially as these models are increasingly deployed in high-stakes and diverse domains, such as healthcare, finance, legal systems, and digital platforms. As predictive systems grow more complex, understanding how and why a model arrives at a particular decision is essential not only for debugging and improvement but also for ensuring fairness, accountability, and user trust. Therefore, developing effective methods to interpret machine learning models is crucial for aligning technical performance with ethical and practical expectations in real-world applications.

Recommendation systems, a key application of machine learning, have become integral in modern digital platforms, connecting users with relevant items across various domains, from e-commerce to entertainment. While traditional machine learning tasks provide precise point predictions, the core objective in recommendation systems is to accurately rank items based on users' predicted preferences. This change in focus underlines the need to adapt explainability techniques to better align with ranking based recommendation systems. Most of the existing explainability methods are effective for explaining individual predictions but they are often insufficient in expressing the comparative logic behind a generated ranked list. For instance, understanding why a model recommends "Item A" over "Item B" is crucial for user trust, system transparency, and even for identifying potential biases.

This paper introduces RanXplain, a methodology specifically designed to address this gap by explaining the comparative behavior of rankings generated by recommendation models. RanXplain functions as an independent machine learning model, trained on pairs of items recommended by the ranking model. It utilizes features derived from the original ranking model, enriched with additional comparison features that capture the differences between items. The application of both inherently explainable models and more complex, high-performing models were explored within RanXplain framework. The approach addresses the unique challenges of explaining rankings, offering flexible and detailed insights into why one item is placed above another in a recommendation list. By doing so, RanXplain aims to increase the transparency and interpretability of recommendation systems, promoting user understanding and trust.

The remainder of the paper is organized as follows: Section II reviews the related work on explainable AI and explanation methods. Section III introduces the RanXplain methodology in detail. Section IV offers the key results and experiments, along with a brief evaluation and discussion. Finally, Section V concludes the paper and outlines directions for future research.

II. RELATED WORK

Explainable Artificial Intelligence (XAI) is now one of the most important topics in many machine learning systems, due to the increasing need for transparency, trustfulness, and accountability [1][2]. With the high adoption of artificial intelligence in various fields, such as healthcare, banking, law, e-commerce, entertainment, interpreting predictions has been as important as creating the predictions themselves. Approaches to XAI may be categorized in terms of their usage with models and explaining the local or global behaviors.

1) Model-Intrinsic (or Inherently Interpretable) vs. Model-Agnostic (or Post-Hoc):
   - Model-intrinsic methods rely on the inherent transparency of certain machine learning algorithms, such as linear models or decision trees, whose internal structures make them naturally suitable for generating explanations.
   - On the other hand, model-agnostic methods are complementary for so-called black box models, such as neural networks, gradient boosting trees in a way that these methods are used after the predictions have been made.

These methods are, therefore, more flexible and may be used with any algorithm.

2) Local vs global explanations:
   - Local explanations aim to clarify individual input-output decisions, such as why a specific application was rejected or why a particular prediction probability was assigned.
   - Global explanations, however, try to give a general image of the behavior of the models and can be thought of as a summary of the model.

Local Interpretable Model-Agnostic Explanations (LIME) [3] and SHapley Additive exPlanations (SHAP) [4] are two popular model-agnostic local explanation approaches designed to explain any given black box classifier. Both of them work as feature attribution linear models, trying to understand the degree of change in predictions and particular features used to generate these predictions.

Even though they are extremely widely used and general, SHAP and similar feature attribution methods are basically limited [5][6][7], especially in ranking tasks. These methods are designed to explain instance-wise predictions by attributing the outcome to each feature one at a time. However, in recommendation systems, where the main task is to rank items relative to one another, such pointwise explanations are not able to capture the relative dynamics among items. For instance, the fact that the particular feature had a positive impact on the score of Item 1 tells us relatively little about the reasons why Item 1 outperformed Item 2. In Figure 1, row-wise SHAP-style feature attributions for the top four recommended items for a user are shown to illustrate this limitation. Each row corresponds to one item, with SHAP values color-coded based on their magnitude and impact within that row. Green indicates positive contribution toward the item's ranking score, and red indicates negative contribution. Though single-item contributions are formulated for each item, they do not provide insight into relative differences that cause the ensuing ranking order. A seemingly logical, yet misleading, approach would be to simply compare feature contributions between two items. For instance, the SHAP value for price feature (Feature 1) could be positive for Item 1 and negative for Item 2. This large, opposing difference in SHAP values might incorrectly suggest that price is the primary reason for the ranking disparity. In reality, Item 2 can be cheaper than Item 1 and other features like user affinity for specific categories (or brands) might be the true drivers, creating these conflicting individual attributions. This means a feature crucial for an item's individual score may be irrelevant when explaining its comparative rank.

Global explanation methods like Permutation Feature Importance [8] or Partial Dependence Plots [9] similarly fall short in explaining the behavior of ranking models. While they can identify influential features on average across predictions, they do not provide specific, contextual information. For example, price is generally the most important factor for ranking models in e-commerce; however, it does not explain why, for a particular user and context, a more expensive Item A might be ranked higher than a cheaper Item B, contrary to average user behavior. These gaps highlight that neither standard local nor global approaches are inherently suited to the comparative nature of ranking explanations, motivating the need for specialized pairwise or listwise approaches.

Item ID	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Score	Rank
1	0.73	0.92	0.56	0.36	0.25	0.11	-0.74	0.58	2.77	1
2	-0.46	0.14	0.10	0.95	-0.04	0.76	0.42	0.04	1.91	2
3	0.51	-0.08	-0.56	0.49	0.84	0.59	-0.24	-0.18	1.37	3
4	0.57	-0.58	-0.71	0.02	0.84	0.77	0.16	0.19	1.27	4

Figure 1. Row-wise SHAP-style feature attributions.

One of the most influential pairwise approaches is the Analytic Hierarchy Process (AHP) and its generalization, the Analytic Network Process (ANP), introduced by Saaty [10][11] for decision-making based on pairwise comparisons. In AHP/ANP, decision-makers explicitly provide judgments on the relative importance of alternatives or criteria, and a priority ranking is then derived using the principal eigenvector of the comparison matrix. This framework has been widely applied in domains, such as project selection, resource allocation, and policy evaluation. The RanXplain framework, however, addresses the inverse problem: instead of deriving rankings from human-provided comparisons, it seeks to explain rankings that have already been produced by machine learning models. While one might envision applying Saaty's eigenvector method directly to model-generated pairwise scores, several practical obstacles arise.

First, the scale of modern recommender systems far exceeds the typical scope of AHP/ANP: a single user session may involve thousands of candidate items (e.g., in e-commerce with catalogs exceeding 10 million products) and hundreds of input features (e.g., user–item embeddings, contextual features, temporal recency signals). Constructing and processing complete $n \times n$ pairwise matrices under such conditions becomes computationally intractable. Second, the eigenvector solution yields overall item priorities but does not provide feature-level contributions to rankings, which are essential for transparency in explainable AI. Third, while AHP assumes relatively stable and consistent comparison judgments, machine-learned rankings are highly context-dependent, with the relative importance of features varying substantially across users and sessions. These distinctions underscore why classical AHP/ANP methods are not directly applicable to explaining large-scale AI ranking systems.

In the following sections, a comparative RanXplain methodology will be discussed in detail on how to mitigate the gaps of the current methods of XAI.

III. Methodology

The methodological framework for the RanXplain model outlined in this section, addresses the aforementioned limitations of existing explainability methods in ranking. RanXplain provides explanations for pairwise preferences within a ranked list of items, clarifying the comparative reasoning of the original ranking model. Effectively, RanXplain operates as a seperate

machine learning model, trained to explain the primary ranking system's comparative behavior.

A. Data Generation for RanXplain: The Pairwise Paradigm

RanXplain focuses on enhancing a personalized recommendation system in terms of explainability. The underlying notion behind RanXplain is to transform the complex problem of explaining the ranking of the whole item list into a series of manageable binary classification problems based on pairwise comparisons. A training instance is constructed for RanXplain, for each relevant pair of items derived from the output of the primary ranking model.

In personalized recommendation systems, the common approach involves generating pointwise predictions for individual user-item pairs. Items are then ranked for each user based on these scores. However, while it might be possible to explain why a single item received a particular prediction score (although even this is often challenging with typical ranking models), it's rarely clear why "Item A is ranked higher than Item B." This explanation is often more intuitive for users trying to understand their preferences.

This lack of clarity regarding relative rankings makes it difficult for both users and developers to grasp the underlying behavior of the recommendation framework. RanXplain addresses this explanatory deficiency by evaluating ranking model behavior through considering combinations of items.

1) Selection of Pairs: RanXplain relies on modeling pairwise preferences to effectively explain the comparative logic of the primary ranking model. However, it is computationally challenging to generate every possible combination from a large set of items. Therefore, a strategic approach to sampling these pairs is vital, not only for practical implementation but also to ensure the most informative pairs of items are included.

The preferred methodology for generating these pairwise comparisons involves two main strategies, both beginning by determining top K items for each user from their recommendation lists.

The first strategy for generating user-item-item indices involves randomly selecting a subset of k ($k < K$) items for each user, from their selected top K recommendations. All possible pairwise combinations are then created from this subset. This ensures that each item within the chosen subset appears in multiple comparisons for that user, providing a substantial set of data for learning specific comparative preferences of the ranking model.

The second strategy initially forms all possible combinations from the entire set of K top items for each user. Then, a random sampling is applied to obtain a comprehensive collection of pairwise comparisons from this potentially vast dataset. This strategy differs from the first as it creates a subset of the original dataset rather than representing the full data. While this can make the model more robust, it has a key drawback: it might miss some pairwise comparisons between items. For example, if we consider three items (i_1, i_2, and i_3) recommended to a user, the first strategy includes all pairwise comparisons (i_1 vs. i_2, i_2 vs. i_3, and i_1 vs. i_3). In contrast, this strategy might include only some of these pairs, which makes it harder to capture three-way (or higher-order) relationships. Furthermore, this approach may introduce greater imbalance in the number of data points per user, which can lead to biased training or decreased generalization performance.

2) Features of RanXplain: Creating meaningful features is crucial for the RanXplain model to learn from and explain the comparative relations. Original feature set $F = \{f_1, f_2, \ldots, f_N\}$ which were used by the primary ranking model to make pointwise predictions are added to the feature set for both items in each pair (i_1, i_2), so that the feature set of RanXplain contains $2N$ item features for each index since both items have N features.

Additionally, a set of comparison features that explicitly capture the relationship between i_1 and i_2 are derived from the features in F. Let x_1 and x_2 be the values of a feature $f_j \in F$ for i_1 and i_2, respectively. A small constant ε (e.g., 10^{-6}) is introduced to handle potential division by zero. Using x_1, x_2, and ε, a set of comparison features F_{comp} is constructed as follows:

Ratio: The ratio of feature values for items i_1 and i_2 is defined as shown in (1):

$$\frac{x_1}{x_2 + \varepsilon} \tag{1}$$

Mean Percentage Error (MPE): The MPE between feature values, as calculated in (2), is computed as:

$$\frac{x_1 - x_2}{x_1 + x_2 + \varepsilon} \tag{2}$$

Difference: The absolute difference between feature values is simply expressed by (3):

$$x_1 - x_2 \tag{3}$$

Relative Deviation: The relative deviation, given by (4), captures the proportional difference:

$$\frac{x_1 - x_2}{x_1 + \varepsilon} \tag{4}$$

Equality Indicator: For categorical features, an indicator function checks equality, as defined in (5):

$$\mathbf{I}_{x_1 = x_2} = \begin{cases} 1 & \text{if } x_1 = x_2, \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The full feature set of RanXplain includes both item features from the original ranking model and features that describe the comparison between item pairs in order for the model to learn more detailed comparison logic. All different combinations of feature sets have been tested by adding and discarding them to optimize the feature set for effective comparative learning.

3) Target Variable: The target variable for RanXplain is a binary indicator, which has the value of 1 if the first item i_1 in the pair (i_1, i_2) is ranked higher by the primary ranking model. If the second item i_2 is ranked higher, the target is 0. This approach turns the primary model's unknown pairwise decisions into a clear, learnable signal for RanXplain.

B. RanXplain Model Selection

Selecting the type of underlying machine learning model is critical for development of RanXplain. It requires balancing robust predictive performance with the need for interpretability when explaining comparative behaviors. Logistic Regression and XGBoost are considered as two prominent models for this purpose.

- Logistic Regression: Initially advanced by [12] and further generalized by [13], Logistic Regression is a linear model which is particularly advantageous for its inherent interpretability in binary classification. RanXplain's aim of interpreting ranking behavior by classifying pairwise preferences, directly aligns with capability of this model type. Within RanXplain, Logistic Regression models the probability that i_1 is prioritized over i_2 by the primary ranking model. Its direct interpretability comes from its learned coefficients:
  - A positive coefficient for a feature $f_j(i_1)$ indicates that an increase in f_j for i_1 directly raises the probability of i_1 being preferred, assuming other features remain constant.
  - Critically, for comparison features, such as $f_j(i_1) - f_j(i_2)$, a positive coefficient directly quantifies that a higher difference in f_j in favor of i_1 contributes proportionally to its higher predicted preference.

  This direct mapping between feature values and their impact on the log-odds of preference provides transparent and comprehensible explanations for the primary ranking model's comparative logic. Its main limitation in this context is its inability to capture complex non-linear relationships or feature interactions that may characterize the primary ranking model's decision-making process.
- XGBoost (Extreme Gradient Boosting): An optimized gradient boosting framework which is introduced by [14], offers superior predictive performance by constructing an ensemble of decision trees. While inherently a black-box model, its utility within RanXplain for generating explanations is realized through the application of SHAP values. SHAP provides a robust, unified framework to attribute the contribution of each feature to a specific prediction.
  - For a RanXplain model trained with XGBoost, SHAP values precisely quantify the impact of each feature on the prediction of whether i_1 is preferred over i_2. This enables local explanations for individual pairs (e.g., attributing i_1's preference to its higher "discount" and "popularity" differential).
  - Furthermore, aggregating SHAP values enables global insights into the most important features effecting comparative preferences across the entire dataset (e.g., identifying "price difference" as a universally strong determinant of higher ranking).

XGBoost's advantage lies in its competency to model complex non-linear relationships and high-order feature interactions, potentially offering a more accurate representation of the primary ranking model's intricate decision boundaries. The need for post-hoc explanation methods like SHAP is the disadvantage of using XGBoost for RanXplain. Although SHAP is a powerful method to produce explanations, it is more complex and computationally intensive than using direct coefficients from Logistic Regression.

C. Explanation Generation and Presentation

The practical applicability of the RanXplain methodology extends beyond its predictive capacity, addressing the non-trivial step of translating its output into useful, understandable explanations for end-users and system designers. This process is fundamentally guided by the ability to use the model's internal feature weights and contributions to pinpoint the most influential factors in a ranking decision.

Consider a real-world e-commerce scenario in which a recommendation system presents a user with a ranked list of products. Within this list, two items are of particular interest: Item A, an expensive shoe from a well-known brand with an applied discount, and Item B, a medium-priced shoe from a common brand without a discount. The primary ranking model prioritizes Item A over Item B, and RanXplain successfully predicts this outcome.

When RanXplain correctly predicts the prioritization of one item over another, its model coefficients (for Logistic Regression) or feature importance values (for XGBoost) reveal which comparison features were most influential. For instance, the model can identify that the difference in discount ratio, relative brand popularity, or the user's affinity for a specific brand were the key drivers behind the ranking. These features, which quantify the relative properties of the two items, allow for the generation of clear and concise explanations.

This capability enables the extraction of concrete insights, such as: "Item A was ranked higher than Item B because, while Item B is cheaper, the model gave more weight to the discount available on Item A and the user's affinity for Item A's brand." This ability to generate detailed, feature-based explanations serves several primary purposes in real-world applications:

- User Trust and Understanding: Providing explanations for why a specific item was prioritized helps users understand the system's logic, leading to increased trust and confidence in the recommendations.
- System Debugging and Improvement: Explanations act as a critical tool for developers to diagnose the primary ranking model's behavior. By analyzing why certain items are ranked in a particular order, developers can identify potential biases, correct model errors, and gain insights for future feature engineering.
- Cross-functional Insights: Explanations can be shared with other teams (e.g., merchandising, marketing) to provide a deeper understanding of customer behavior and content performance. For example, by analyzing explanations, a merchandising team could determine that a 10% price decrease on a specific product would cause it to be ranked higher than a competitor's product for a particular segment of users.

IV. RESULTS | EVALUATION

Experimental evaluation of RanXplain involves a rigorous process, beginning with the detailed construction of three distinct datasets: (i) training set, (ii) test set and (iii) simulation set. The training dataset was formed using top 50 recommendations per user generated by the ranking model within a specified historical period. It consists of 4.5 million rows by over 30,000 unique users and more than 130,000 distinct items while maintaining a balanced 50% target ratio. As the test set has been obtained by splitting the initial training set according to 80%-20% parity, it contains 1.1 million rows while exhibiting comparable unique user and item counts and the same 50% target ratio. Crucially, simulation dataset, consisting of 50 million rows, was generated by incorporating all 50 top-ranked items for each user from a later temporal period than the training set, including approximately 45,000 users and over 175,000 distinct items, also with a 50% target ratio.

The choice of sampling strategy is crucial for both the practical impact and computational efficiency of RanXplain. By transforming the complex task of explaining a ranked list into a series of pairwise classification problems, RanXplain becomes computationally tractable for large-scale recommendation systems, which is a significant advantage over other methods. Two different sampling strategies were explored for training RanXplain: (i) content-based sampling and (ii) random sampling. Performance metrics of the models trained on both datasets were observed as highly similar. However, content-based sampling yielded slightly superior performance and provided a more representative distribution across diverse users. This intentional sampling approach makes the training process more efficient and ensures that the resulting explanations are representative and of high quality, which is vital for real-world application. Consequently, content-based sampling method was adopted by randomly selecting 20 items per user from their top 50 recommendations and generating all possible combinations for the selected 20 items.

Experiments of RanXplain proceeded to exploring two critical dimensions in more detail: feature set composition and model architecture, concluding with a detailed simulation-based performance evaluation.

A. Experimentation of Feature Sets

To investigate the impact of features on RanXplain model, a set of experiments were conducted. Table I depicts performance metrics across train, test and simulation datasets of the Logistic Regression models trained with different feature sets in BigQuery ML [15]. Performance of the models were assessed using the Receiver Operating Characteristic Area Under the Curve (ROC-AUC), which quantifies the ability of a classifier to discriminate between positive and negative classes across various thresholds [16]. Similar behavior was observed across other performance metrics, such as accuracy and recall.

Initially, Model 1 was trained using only item features which is an approach that mirrored the original ranking model. By incorporating comparison features alongside these item features in Model 2, a significant improvement in model performance was observed across all ROC-AUC metrics for the training, test, and simulation datasets.

TABLE I. RANXPLAIN MODEL PERFORMANCE WITH DIFFERENT FEATURE SETS

Metric	Model 1	Model 2	Model 3
Item Features	Included	Included	Excluded
Comparison Features	Excluded	Included	Included
Train ROC-AUC	0.62	0.73	0.74
Test ROC-AUC	0.61	0.74	0.74
Simulation ROC-AUC	0.61	0.69	0.70

Interestingly, Model 3 which is trained exclusively with comparison features achieved slightly better predictive performance than the models with item features. While the predictive gains were marginal, using only comparison features significantly enhanced the qualitative aspect of explanations compared to Model 2. The increase in qualitative aspect is due to the directness of interpretability that comparison features provide when comparing two items.

Slightly improved performance along with stronger interpretability indicates the vital role of comparison features in accurately capturing the relative ranking of items. Therefore, the comparison features are adopted as the feature set of RanXplain.

B. Experimentation of Model Types

For model selection, Table II shows performances of models that differ by model type and maximum tree depth. Although Model 3 was the best performer in the experiments of feature sets, Model 2 was chosen as a baseline model to be compared with XGBoost models (which are trained using BigQuery ML [17]) so that both item and comparison features are included in experimentation of model types.

TABLE II. RANXPLAIN PERFORMANCE FOR DIFFERENT MODELS

Metric	Model 2	Model 4	Model 5
Model Type	Log Reg	XGBoost	XGBoost
Max Tree Depth	-	15	5
Train ROC-AUC	0.73	0.92	0.79
Test ROC-AUC	0.74	0.92	0.79
Simulation ROC-AUC	0.69	0.78	0.69

Reducing the maximum tree depth in the XGBoost model causes significant decrease in model performance across all train, test and simulation sets. This decrease is evident in Table II, as shown by the performance difference between Model 4 and Model 5. This observation motivates the use of a more sophisticated XGBoost model within RanXplain. Additional complexity is required to effectively approximate the behavior of primary ranking model, which is a highly complex model. However, while higher complexity increases the prediction performance, it also makes the interpretation of the explanation model more challenging.

It can be inferred from Table II that Model 2 underperformed Model 4 with respect to ROC-AUC metrics. However, Logistic Regression possesses inherent interpretability as opposed to XGBoost which requires additional methods like SHAP for explanations. Although more complex models like XGBoost might be a better fit depending on the specific application's requirements, Logistic Regression is found more suitable for RanXplain of the primary ranking model that is used in this study.

C. Simulation Based Performance Evaluation

Simulation dataset was used to conduct various offline evaluations on the preferred model (Model 3). Consistent ROC-AUC performance was observed across the training, test, and simulation datasets, with only a slight performance decrease on the simulation set. Further analyses were conducted on the simulation data to understand the behavior of RanXplain model more comprehensively. These analyses are based on two key factors: (i) proximity of item ranks in the original ranking model and (ii) whether item pairs belonged to the same high level item category (e.g., electronics category).

Table III depicts train, test and simulation performances of Model 3. Simulation performance was analysed with respect to three additional breakdowns: subset of the simulation data (i) where the difference between rankings of two items are greater than 20 ($r_{diff} > 20$) (ii) where the difference between rankings of two items are less than or equal to 3 ($r_{diff} \leq 3$) and (iii) where the two items belong to the same category (Same Category).

TABLE III. SIMULATION PERFORMANCE

Metric	Model 3
Train ROC-AUC	0.74
Test ROC-AUC	0.74
Simulation ROC-AUC	0.70
Simulation ROC-AUC ($r_{diff} > 20$)	0.80
Simulation ROC-AUC ($r_{diff} \leq 3$)	0.54
Simulation ROC-AUC (Same Category)	0.70

The results revealed a clear trend, predictive capability of the model significantly improves as the rank difference between item pairs increases. For example, the model performed substantially better when the rank difference exceeded 20, achieving an ROC-AUC of 0.80. On the contrary, performance dropped considerably for closely ranked items (($r_{diff} \leq 3$)), with an ROC-AUC of approximately 0.54. This finding indicates that RanXplain has difficulty in predicting (and thus explaining) prioritization when the primary ranking model assigns similar scores to items, which is expected.

This behavior is a key advantage of the RanXplain approach, as it allows us to know in advance when its outputs can be used to confidently interpret the ranking model's decisions, thereby preventing misleading or false insights. The correctness of RanXplain's predictions (and therefore their reliability for generating insights) is known in advance, since the real rankings are already known. This enables the clear identification of when it is safe to use RanXplain's outputs to interpret the behavior of the underlying ranking model for specific item pairs, thereby avoiding misleading or false insights.

Regarding category influence, RanXplain's predictions for pairs within the same item category were very similar to its performance on pairs from the whole simulation set, indicating no significant performance differential. Consequently, for the application of this study, improving RanXplain's performance on closely ranked pairs is of minor importance, although such improvements are feasible by adjusting sampling strategies or incorporating additional comparison features.

V. Conclusion and Future Work

This paper introduced RanXplain, a methodology designed to address a significant gap in recommendation systems, which is the need to explain ranking decisions rather than individual item predictions. As outlined in the previous sections, RanXplain functions as a seperate machine learning model trained on item pairs, employs features derived from the original ranking model. Both the effectiveness and operational behavior of RanXplain is illustrated through a systematic investigation of various feature sets and model architectures, along with simulation-based performance evaluation.

The main contribution of RanXplain lies in shifting the focus of explainability from pointwise predictions to the comparative logic behind ranked outputs. RanXplain enables a more intuitive and actionable understanding of why one item is ranked above another by reframing the task of explaining a ranked list as a series of pairwise classification problems. The aim is to provide interpretable insights into the decision-making process of black-box recommendation models, supporting user trust and contributing to system debugging.

The evaluation based on ROC-AUC across various datasets highlighted the strong influence of comparison features. Models trained exclusively on these features not only achieved better predictive performance but also yielded more interpretable explanations as a result of the direct relevance of the input features. While more complex models, such as XGBoost, offered better predictive performance, Logistic Regression proved to be more suitable for applications that require interpretability, even at a modest cost to accuracy.

The simulation based evaluation further revealed that RanXplain's predictive performance improves significantly as the rank difference between items increases. On the other hand, its performance naturally decreased when items were very closely ranked, which is expected given that the ranking model assigns similar scores in such cases. It is important to note that one of RanXplain's primary advantages is that the correctness of its predictions is known in advance, since the ground truth rankings are available. This capability allows for the identification of cases where RanXplain's outputs can be confidently used to interpret the ranking model's decisions, thus avoiding potential misinterpretations. This observation also draws a parallel to the concept of rank reversal in pairwise

comparison methods like Saaty's AHP, suggesting that the underlying ranking decisions for closely-ranked items are inherently more ambiguous and less stable, making them difficult to explain with high confidence.

The approach shows promise in interpreting pairwise relative rankings; however, RanXplain is not designed to provide a single, holistic explanation for an entire ranked list. While a high-level explanation might be desirable, it can often be too simplistic to capture the nuanced decision-making process of a complex ranking model. Instead, RanXplain provides a series of granular, actionable insights. An explanation for an entire ranked list can be composed by chaining together a series of pairwise comparisons, such as explaining why Item 1 was ranked above Item 2, why Item 2 was ranked above Item 3, and so on. This approach offers a more detailed and accurate understanding of the ranking process, as it clearly articulates the specific feature-level trade-offs that led to the final ordering. This modular nature allows RanXplain to provide highly specific insights on demand, supporting both user understanding and system debugging by clarifying the reasons behind individual ranking decisions.

For future work, several promising directions can be explored to further enhance RanXplain. The AUC performance of the model, particularly on closely ranked pairs, can be enhanced through various methods. This could involve incorporating additional non-linear comparison features, such as the power of the difference of feature values, to better capture the primary model's complex decision boundaries. Furthermore, exploring alternative and more advanced sampling techniques or using a wider range of training data could lead to significant improvements in model performance and a more robust understanding of the ranking model's behavior. An extension of RanXplain to support counterfactual explanations could offer more actionable insights for users and system designers by indicating how changes in specific features would affect the relative ranking of items. The trade-off between user-based and random sampling, and how different sampling strategies impact the quality of explanations, presents a key area for further research. RanXplain can also be used in a reverse engineering context to guide feature design in the original ranking model. When important comparison features are identified but prove insufficient on their own, new supporting features can be introduced to the original ranking model. This can improve both the model's performance and its explainability.

ACKNOWLEDGMENT

This study was supported by Trendyol's R&D Center through infrastructure and organizational contributions. The work was carried out within the scope of an R&D project approved by the Republic of Türkiye Ministry of Industry and Technology.

REFERENCES

[1] V. Hassija et al., "Interpreting black-box models: A review on explainable artificial intelligence", *Cognitive Computation*, vol. 16, 2023. DOI: 10.1007/s12559-023-10179-8

[2] Y. Zhang and X. Chen, "Explainable recommendation: A survey and new perspectives", *Foundations and Trends in Information Retrieval*, vol. 14, no. 1, pp. 1–101, Mar. 2020. DOI: 10.1561/1500000066

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, New York, NY, USA: Association for Computing Machinery, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions", in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)*, Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 4768–4777.

[5] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling lime and shap: Adversarial attacks on post hoc explanation methods", in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ACM, 2020, pp. 180–186. DOI: 10.1145/3375627.3375830

[6] S. Bordt, M. Finck, E. Raidl, and U. von Luxburg, "Post-hoc explanations fail to achieve their purpose in adversarial contexts", in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, New York, NY, USA: Association for Computing Machinery, 2022, pp. 891–905. DOI: 10.1145/3531146.3533153

[7] H. Lakkaraju and O. Bastani, ""how do i fool you?": Manipulating user trust via misleading black box explanations", in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*, New York, NY, USA: Association for Computing Machinery, 2020, pp. 79–85. DOI: 10.1145/3375627.3375833

[8] J. H. Friedman, "Greedy function approximation: A gradient boosting machine", *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[9] L. Breiman, "Random forests", *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[10] T. L. Saaty, *The Analytic Hierarchy Process*. McGraw-Hill, 1980.

[11] T. L. Saaty, *Decision Making with Dependence and Feedback: The Analytic Network Process*. RWS Publications, 1996.

[12] J. Berkson, "Application of the logistic function to bio-assay", *Journal of the American Statistical Association*, vol. 39, no. 227, pp. 357–365, 1944.

[13] D. R. Cox, "The regression analysis of binary sequences", *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 20, no. 2, pp. 215–242, 1958.

[14] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, San Francisco, CA, USA: ACM, Aug. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785

[15] Google Cloud, "Create model statement for logistic regression models (glm)", Accessed: 2025.07.18, 2024. [Online]. Available: https://cloud.google.com/bigquery/docs/reference/standard-sql/bigqueryml-syntax-create-glm

[16] T. Fawcett, "An introduction to roc analysis", *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006. DOI: 10.1016/j.patrec.2005.10.010

[17] Google Cloud, "Create model statement for boosted tree models (boosted_tree)", Accessed: 2025.07.18, 2024. [Online]. Available: https://cloud.google.com/bigquery/docs/reference/standard-sql/bigqueryml-syntax-create-boosted-tree

Bridging the Domain Gap: Evaluating Fact-Grounded Knowledge Graph Narratives for Explainable AI in Clinical Decision Support

Valentin Göttisheim, Holger Ziekow,
Peter Schanbacher
Faculty of Business Information Systems
Furtwangen University
Furtwangen, Germany
e-mail: firstname.surename@hs-furtwangen.de

Djaffar Ould-Abdelsam
Université de Haute-Alsace
IRIMAS Laboratory, Université de Haute-Alsace
68100 Mulhouse, France
e-mail: djaffar.ould-Abdelsam@uha.fr

***Abstract*—Clinicians need transparent reasoning to trust Artificial Intelligence recommendations, but standard explanation methods lack clinical semantics. To address this, we transform an Onkopedia colon carcinoma guideline into a semantically enriched Knowledge Graph by segmenting text, extracting and merging semantic concepts, enriching gaps with registry data, and anchoring features to graph nodes. Using a predictive model, we compute Shapley Additive Explanations feature attributions and generate fact-grounded narratives via large language models that directly reference guideline evidence. We compare three contexts across 65 synthetic colorectal cancer cases (195 narratives) and find that KG-based narratives reduce hallucinations, speculation, and contradictions. Embedding KG-grounded narratives in clinical decision-support tools promises to shorten expert review cycles, surface guideline deviations, and bridge the explainability gap between data scientists and clinicians.**

Keywords-Keywords— Explainable Artificial Intelligence; XAI; Knowledge Graphs; Shapley Additive Explanations; SHAP; Narrative Generation; Claim Verification.

I. Introduction

Clinical decision support models promise early insights but often function as opaque black boxes [1]. Clinicians require transparent, evidence-based explanations to understand how input features drive predictions [2]. In practice, model development is a collaborative, iterative process: data scientists train and refine predictive models, generate interim explanations, and oncologists review these artifacts against clinical knowledge, suggest adjustments, and feed feedback into retraining until statistical performance and clinical relevance converge. This real-world feedback loop motivates our work.

To bridge the gap between raw model outputs and clinically meaningful interpretation, we augment Shapley Additive Explanations (SHAP) outputs with fact-grounded narratives linked to an authoritative guideline-derived Knowledge Graph (KG). Our contributions are threefold:

1) Extract and structure clinical guideline content into a semantically rich KG.
2) Compute SHAP attributions for model features and anchor them to KG nodes.
3) Generate narrative explanations referencing the KG, yielding traceable, domain-specific rationales.

Standard SHAP bar charts quantify feature influence but lack clinical semantics. By mapping attributions to KG nodes derived from colon carcinoma guidelines, our approach enriches explanations with medical context—enabling clinicians to reason in domain-specific terms and data scientists to identify discrepancies from accepted evidence. We therefore ask how such fact-grounded narratives affect four claim categories—*Hallucination*, *Contradiction*, *Speculation*, and *Extrapolation*:

(RQ1) Does KG anchoring reduce hallucinations?
(RQ2) Does KG anchoring reduce contradictions?
(RQ3) Does KG anchoring reduce speculative statements?
(RQ4) Does it keep extrapolations within the boundaries established by using guideline text alone?

If successful, this strategy could streamline expert review and facilitate the way for prospective clinical validation. The remainder of the paper is organized as follows: In Section III we present the proposed methods, including KG construction and narrative generation. Section IV reports quantitative and qualitative results. Section V discusses implications and limitations. Section VI concludes with future directions.

II. Related Work

Shapley values provide theoretically grounded, local feature attributions that have become standard in explainable clinical ML [3], but dense bar-chart displays impose high cognitive load on physicians [4]. To improve interpretability, template-based systems, such as *SHAPstories*, convert attributions into short rationales, yielding modest trust gains [5], while constrained decoding in *EXPLINGO* reduces hallucinations in general domains [6]. Burton et al. frame explanation verbalization as a data-to-text task with the TEXEN corpus—496 SHAP/LIME-to-narrative pairs—reporting factual error rates of 25%-42% for models like BART and T5 [7]. Although these methods enhance usability, they lack integration with domain-specific clinical knowledge.

Evaluation of explanation quality typically distinguishes between *faithfulness*—how accurately an explanation reflects the underlying model—and *plausibility*—how well it aligns with human judgment [8–10]. Kroeger et al. demonstrate that larger language models can yield less faithful post-hoc explanations without additional constraints [11], and Lanham

et al. offer a fine-grained benchmark for faithfulness in chain-of-thought reasoning [12]. Diagnostic probes, such as Walk-the-Talk and the FaithEval suite, complement traditional lexical overlap metrics (BLEU, ROUGE) by assessing deeper semantic and factual fidelity [13][14]. To build upon this strand, we introduce a structured factual-consistency framework that quantifies divergences across four categories: *Hallucination*, *Extrapolation*, *Speculation*, and *Contradiction*, as defined in Table II and applied in Table IV.

Knowledge Graphs enhance semantic structure, traceability, and bias control in otherwise opaque model explanations [15]. Typical KG construction pipelines involve text segmentation, entity and relation extraction, canonicalization, ontology alignment, and population [15], while widely used biomedical resources, like the UMLS Metathesaurus and Bio2RDF, integrate millions of curated concepts from diverse ontologies [16]. Domain-grounding systems, such as *XplainLLM*, anchor generated explanations in KG triples; *DR.KNOWS* integrates UMLS—a large compendium of biomedical terminologies—for diagnostic safety [17][18]. Cross-domain cybersecurity work highlights that LLM-based verbalization of SHAP tables can still wander off-fact without authoritative grounding [19]. Emerging LLM-based tools (e.g., Text2KG, LLM-Assisted Knowledge Graph Engineering) automate parts of these pipelines but face challenges, such as hallucination and schema drift [20][21]. Crucially, no existing approach constructs KGs directly from prescriptive clinical guidelines—a gap our guideline-driven pipeline addresses by extracting semantic concepts from Onkopedia guidelines, enriching them with registry data, and anchoring model features to KG nodes.

Building on post-hoc feature attributions (SHAP), narrative verbalization, domain-specific evaluation metrics, and established KG construction pipelines, we address the challenge of grounding model explanations in clinical evidence. We integrate guideline-derived Knowledge Graph construction with SHAP-anchored narrative generation to produce explanations that are both interpretable and verifiable. We evaluate factual accuracy by fact-checking statements in the generated narratives against patient case records and quantify divergences from the ground truth. This methodology yields fact-anchored narratives that clinicians can immediately verify against clinical guidelines, enhancing trust and accelerating prospective validation.

III. Proposed Methods

We developed an end-to-end pipeline that (i) transforms the Onkopedia colorectal-cancer (CRC) guideline [22] into a semantically enriched Knowledge Graph, (ii) computes Shapley Additive Explanations attributions on an XGBoost predictive model to quantify feature importance, and (iii) generates fact-grounded narrative explanations via large language models (LLMs), which we evaluate experimentally for factual consistency.

A. Knowledge Graph Representation

We represent the guideline-derived KG as a labeled directed graph, where nodes correspond to clinical semantic concepts (e.g., therapies, biomarkers, patient characteristics), edges denote typed relationships between them, and both nodes and edges carry labels derived from the medical guideline.

We implemented a six-stage pipeline to transform the CRC guideline into a semantically enriched Knowledge Graph:

Step 1: Preprocessing & Chunking: Clean raw guideline text (remove headers, footers) and segment into traceable 100-character chunks with metadata (chapter, page, hash).

Step 2: Concept & Relation Extraction: Apply GPT-o4-mini-high with structured prompts to extract semantic concepts as entities with attributes (name, description, confidence) and their inter-relations into a validated JSON schema.

Step 3: Subgraph Integration & Clustering: Merge chunk-level subgraphs into an initial graph, cluster entities by thematic category, consolidate identical identifiers, and link synonyms.

Step 4: Registry Enrichment: Identify missing clinical concepts, insert placeholder nodes, and enrich them with real-world CRC registry attributes (e.g., age, KRAS status, ECOG).

Step 5: Master Graph Assembly: Integrate all enriched subgraphs under a central root node, serialize in Markdown, and export to Neo4j format for queryability.

Step 6: Provenance Annotation: Attach detailed source metadata (document, chapter, page, chunk ID, hash) to every node and edge for auditability.

B. Narrative Generation

Based on a real-world colorectal-cancer registry data schema excerpt provided by our research partner, we built a simulation and generated 20,000 synthetic patient records. We trained an XGBoost model to forecast patient-level treatment decisions and quantified feature importance with SHAP contribution scores (ϕ_i) using the TreeExplainer algorithm [3]. SHAP decomposes each prediction $f(\mathbf{x})$ as:

$$f(\mathbf{x}) = \phi_0 + \sum_{i=1}^{M} \phi_i,$$

where ϕ_0 is the model's expected output and each ϕ_i the marginal contribution of feature i. We linked features to their corresponding nodes in the guideline-derived KG, ensuring semantic grounding. However, not all features can be anchored to the KG, since some registry variables (e.g., body mass index or weeks since initial diagnosis) are not guideline-based clinical concepts. We then synthesized 65 colorectal-cancer patient personas—each defined by demographic variables, TNM stage, ECOG performance status, Charlson Comorbidity Index [23], and molecular biomarker profile—and stratified them into three complexity tiers: (i) uncomplicated cases without guideline conflicts; (ii) biomarker-driven cases; and (iii) multimorbid cases with conflicting recommendations. For each persona, we computed SHAP attributions using TreeExplainer on the

XGBoost predictive model and selected the ten highest-impact features by absolute SHAP magnitude. We then generated narrative explanations in three grounding contexts (OA, GL, KG), defined in Table I, with GPT-o4-mini-high, supplying both the complete patient CSV record and the top-ten SHAP features as patient case data. This 3 × 65 factorial design produced 195 narratives, enabling paired comparisons of factual consistency across grounding strategies. To evaluate the incremental impact of integrating clinical guidelines and Knowledge Graph information, we prompt the LLM (GPT-o4-mini-high) to generate narrative explanations under the three controlled contexts (OA, GL, KG). All narratives follow a standardized Markdown template to control for length and format, ensuring identical format and length constraints across experimental conditions.

TABLE I. GROUNDING CONTEXTS FOR NARRATIVE GENERATION

Context	Description
OA (Only-Attributes)	Patient case data alone, excluding guideline or KG context.
GL (Guideline)	Patient case data plus extracted guideline excerpts with explicit citations.
KG (Knowledge Graph)	Patient case data and full KG in Markdown, including labels, relations, and provenance.

C. Claim Extraction and Evidence Matching

We parsed each created narrative with GPT-o4-mini-high to extract individual asserted claims (complete sentences). For each claim, we matched its content against the patient case data (patient attributes and corresponding SHAP attributions). The LLM was prompted to flag each claim without direct support in the patient case data as *inferred* and to classify it into four categories: **Hallucination**, **Contradiction**, **Extrapolation**, and **Speculation**, as defined in Table II.

TABLE II. INFERRED CLAIM CATEGORIES AND DEFINITIONS

Category	Why the claim is inferred
Hallucination	The claim asserts a patient-specific fact that is *not present* in the case data or SHAP features; the model introduces new clinical information not observed in the input.
Contradiction	Claim **conflicts** with patient case data.
Extrapolation	Guideline-consistent generalization that lacks direct case evidence.
Speculation	Conjecture with insufficient grounding (not verifiable against case or guideline).

In the following, we illustrate examples of the LLM evaluated claim extraction and evidence matching phase. Each category in Table II is exemplified with excerpts from the LLM evaluation to illustrate the four distinct ways in which a generated *inferred* claim can arise. According to the **Extrapolation** criterion, a claim is clinically plausible and drawn from the guideline but lacks direct support in the patient record. For example:

> *"For a patient with stage I (T2 N0 M0) colon carcinoma, complete surgical resection is curative and no adjuvant chemotherapy is indicated."*

Here, the tumor stage (T2 N0 M0) is correctly taken from the case data, yet the recommendation about cure and omission of chemotherapy, while guideline-based, cannot be verified against any patient-specific attribute. Such extrapolations are nevertheless desirable, because they showcase the language model's ability to enrich its output with domain knowledge and provide broader narrative explanations rather than relying solely on SHAP-derived feature attributions. A **Speculation** covers plausible inferences that nonetheless lack explicit evidence. For example:

> *"ECOG 1 (–0.12) and a high comorbidity burden (CDRRHIGH_yes, –0.10) further lowered the probability because of toxicity concerns."*

Although ECOG and comorbidity are real features, attributing the SHAP-driven probability drop to "toxicity concerns" is conjectural and not encoded in the patient case. Such speculation are undesirable, as it introduces clinical reasoning not backed by case data and can mislead users about the true factors influencing the model. By contrast, a **Hallucination** arises when the model fabricates a patient-specific fact that does not appear in the input at all. Consider:

> *"Difference 1: According to the guidelines, an anti-EGFR antibody should be added for RAS-wild-type disease, whereas the model instead selects a BRAF-targeted agent (AB)."*

This statement wrongly attributes BRAF targeting to AB—a fact not mentioned in the case data. Such hallucinations are undesirable because they introduce clinical assertions not backed by case data, undermining trust in the explanation and potentially misleading downstream decisions.

Finally, **Contradiction** occurs when a claim directly conflicts with documented attributes. For instance:

> *"This 55-year-old man with resected rectal cancer (T3 N1 M1) and solitary liver and lung metastases has undergone complete surgical removal of all metastases."*

This contradicts the record's single-metastasis count (`NUMBER_METASTASES`=1) and notes R0 resection only for the primary tumor. Such contradictions are undesirable because they misrepresent case facts.

To validate claim extraction and evidence matching, which were performed automatically using the OpenAI GPT-o4-mini-high model, we randomly sampled 20 claims and computed classification accuracy with 95% Wilson-score confidence intervals to account for small-sample inference [24]. The LLM correctly classified 19 out of 20 cases (95% accuracy), yielding a Wilson 95% confidence interval of 76.4%–99.1%. Even at the lower bound, fewer than 25% of labels are expected to

be incorrect, justifying the use of automatic evaluation for the quantitative analyses.

IV. Results

To evaluate the factual consistency of the generated narratives across three grounding contexts—KG, GL, and OA—we report both quantitative counts and qualitative examples. Results are presented in three parts: overall observed vs. inferred claim counts, composition of inferred categories, and evaluation reliability.

A. Observed vs. Inferred Claims

We evaluated the generated narratives and labeled every asserted claim as either *observed* or *inferred*. A claim is *observed* when it is directly supported by the patient record (e.g., tumor stage or biomarker status) or explicitly grounded by a SHAP attribution that links a named feature to the model's prediction. A claim is *inferred* when it lacks such direct support; inferred claims were further categorized.

TABLE III. Overall observed vs. inferred claim counts by context

Context	Total	✓	❍	% Observed
KG	1 128	367	761	32.5 %
GL	1 125	243	882	21.6 %
OA	1 107	395	712	35.7 %

We report the proportion of observed versus inferred claims across the 195 narratives. Table III summarizes the total number of observed (✓) and inferred (❍) claims across the three grounding contexts. Narratives generated with KG grounding achieved 32.5 % observed claims (367/1 128), outperforming the GL context, which yielded only 21.6 % (243/1 125). The OA context performed comparably to KG with 35.7 % observed claims (395/1 107 vs. KG).

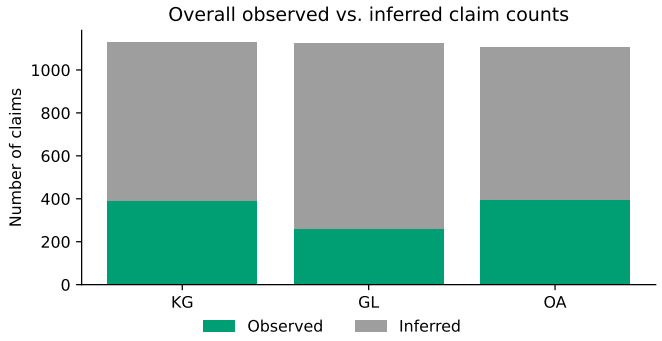

Figure 1. Overall observed vs. inferred claim counts by context (observed = case/SHAP-backed; inferred = not directly case-backed).

Figure 1 plots overall observed vs. inferred claim counts by context. Observed shares differed across the three contexts: explanations grounded in the KG achieved higher observed shares than those from the GL baseline, while OA and KG did not differ much. These findings indicate that KG-grounded input improves consistency over GL-context narratives, while OA may benefits from a narrower input scope with fewer opportunities for *inferred* claims.

B. Inferred Claim Categories

Table IV details the distribution of *inferred* claims by category—**Extrapolation**, **Speculation**, **Hallucination**, and **Contradiction**—expressed as a percentage of total claims in each context.

TABLE IV. Inferred claim category rates (percentage of total claims)

Category	KG	GL	OA
Extrapolation	64.8 %	73.7 %	61.9 %
Speculation	0.5 %	2.0 %	1.1 %
Hallucination	0.0 %	0.4 %	0.2 %
Contradiction	0.1 %	0.6 %	1.1 %

Extrapolation is the predominant inferred category across all contexts. However, the KG condition achieves substantial gains in factual precision and safety: no hallucinations were observed under this setup (0.0 %), speculation drops to 0.5 %, and contradictions fall to just 0.1 %. In contrast, the GL context shows higher rates of speculation (2.0 %) and contradiction (0.6 %).

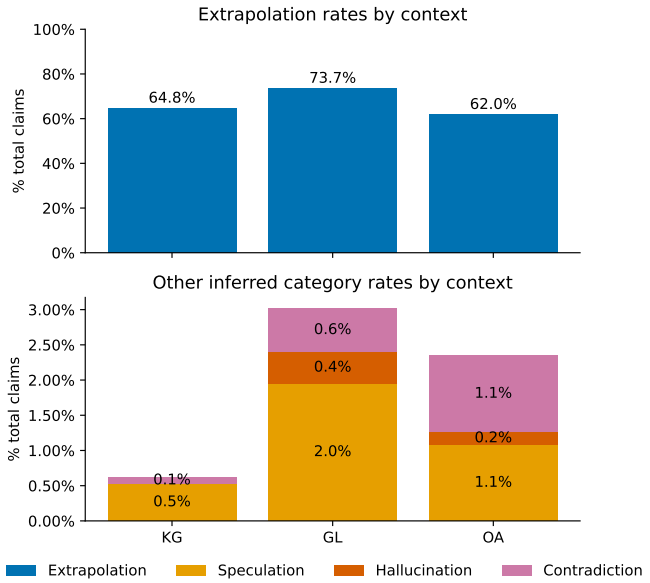

Figure 2. Inferred claim category composition per context (% of total claims).

Figure 2 visualizes these differences as stacked bars (% of total claims). The KG approach yields markedly fewer speculative and contradictory issues than both the GL and OA baselines, and reduces extrapolation by nine percentage points compared to GL. Despite these gains, many claims remain *inferred*, reflecting our design choice to allow clinically plausible, guideline-based extrapolations that may not be explicitly present in the patient record. These results support RQ1 (hallucination), RQ2 (contradiction), RQ3 (speculation), and RQ4 (extrapolation).

C. Qualitative Illustrations

Table V presents an excerpt of one narrative of the same case under the different grounding contexts. The *KG* narrative cites a unique guideline node *[27205d9]* and the recorded feature *RAS|wildtype*, both verifiable in the case file, demonstrating domain-rich yet fact-bound explanation. By comparison, the *GL* narrative, while fluent, infers "stage III disease" solely from *N1* and offers no patient-specific evidence for adjuvant need, showing readability at the expense of precision. The *OA* excerpt repeats guideline buzzwords ("high-risk stage III") relying on generic statements (*T3 N1 M0*), resulting in the most vague prose. For completeness, the last example in Table V presents an GL hallucination example. The mentioned fact—*"left-sided tumor (+0.04)"*—illustrates a feature not present in the patient case and most likely misattributed from the referenced guideline's (*§6.1.4.3.1.1*) metastatic EGFR-therapy discussion, underscoring how lack of authoritative grounding can introduce factual errors.

TABLE V. REPRESENTATIVE NARRATIVE EXCERPTS ACROSS GROUNDING CONTEXTS, WITH GL HALLUCINATION EXPLICITLY MARKED

Context	Narrative Excerpt
KG	Both guideline and model utilise an oxaliplatin + fluoropyrimidine backbone *[27205d9]*; the SHAP feature *RAS|wildtype* supports full cytotoxic sensitivity.
GL	The SHAP value for *N1 (0.28)* flags stage III disease and confirms the need for adjuvant therapy *(guideline §6.1.3)*.
OA	Both guideline and model emphasise high-risk stage III features *(T3 N1 M0)* as key drivers of therapy intensification.
GL	**Hallucination:** RAS wildtype (+0.03) and *left-sided tumor (+0.04)* slightly increased probability, mapping to metastatic guidelines for EGFR-directed therapy *(guideline §6.1.4.3.1.1)*.

The GL hallucination example highlights a reference to a non-existent feature (*left-sided tumor*).

Together, these qualitative vignettes also reinforce our quantitative results: The KG-grounded narrative delivers deep, context-rich explanations that remain verifiable, while the GL outputs sacrifice fidelity for readability and the OA outputs rely on overly generic statements, evidencing a tendency toward vagueness.

V. DISCUSSION

Our study demonstrates that anchoring narrative explanations in a guideline-derived KG improves factual reliability. The KG context reduced hallucinations to 0.0% of total claims in our sample—i.e., none were observed under this setup—supporting RQ1. Moreover, contradictions dropped to 0.1% and speculative claims to 0.5% of total claims, supporting RQ2 and RQ3 that KG grounding reduces both contradictions and speculation.

Moreover, anchoring explanations in the KG cut extrapolation rates from 73.7 % under the GL context to 64.8 %—a 9.0 percentage-point drop—demonstrating that guideline-derived KG grounding effectively constrains extrapolations to within established bounds and thereby confirms RQ4 (See Table IV).

Although the OA context exceeds KG in overall observed-claim rate (35.7% vs. 32.5%), its narrower input scope yields shallower, less semantically rich narratives. OA's lower extrapolation rate (61.9%) comes at the expense of actionable detail, whereas KG grounding delivers fully audit-ready, guideline-anchored explanations (See Table III and Figure 2). Finally, the relatively high share of *inferred* claims across conditions largely reflects clinically plausible, guideline-based extrapolations that provide useful framing but may not be directly present in patient records. In settings that require stricter evidencing, prompts or decoding constraints can restrict extrapolation at the cost of brevity; conversely, future work may calibrate this trade-off per user role (e.g., clinical vs. data science review).

These findings extend prior LLM explainers by showing that structured KG context not only enriches inference but also constrains factual drift [7]. We note that the absence of hallucinations should not be interpreted as impossibility; rather, it likely reflects the combination of KG constraints and the controlled, synthetic case distribution used here.

In practice, clinicians must rapidly validate AI recommendations. The traceable paths in KG narratives—linking each feature attribution to specific guideline nodes—can reduce expert review time by directly surfacing conflicts or affirmations in the guideline text. In our qualitative examples (Table V), KG narratives allowed unambiguous verification of treatment rationale, whereas GL outputs required additional cross-checking. We anticipate that integrating KG-grounded narratives into decision-support dashboards will shorten iteration cycles between data scientists and clinicians, as envisaged in collaborative AI workflows [25].

Our evaluation is constrained by some factors. First, we used 65 synthetic patient personas rather than real-world cases; while this allowed controlled variation, it may not capture the full complexity of clinical data. Second, we benchmarked against a single guideline (Onkopedia CRC) and one LLM version (GPT-o4-mini-high); generalization to other specialties or model variants remains to be demonstrated. Third, our error annotations—though 95% accurate in spot-checks—rely on an automated evaluation LLM; residual misclassifications could slightly bias absolute error rates. Finally, we measured only claim-level errors; additional dimensions such as usability, cognitive load, and end-user satisfaction were not assessed here.

VI. CONCLUSION AND FUTURE WORK

Having demonstrated through our evaluations that KG-grounded narrative explanations outperform both attribute-only and guideline-excerpt baselines in factual reliability, we now outline directions to build on this work. To address limitations and extend our findings, we propose the following directions: (1) Apply the pipeline to real-world data and diverse guidelines;

quantify clinician review time and simulated decision impact. (2) Iteratively refine KG-narrative prompts with user feedback and on-the-fly graph augmentation, aligning with human-centered XAI [5]. (3) Evaluate usability, trust calibration, and clinical actionability; extend metrics (e.g., comprehensiveness, empowerment).

Overall, our results confirm that fact-grounded narrative explanations built on guideline-derived Knowledge Graphs deliver superior factual reliability and coherence compared to attribute-only or guideline-excerpt baselines. By transparently linking model attributions to clinical evidence, this approach paves the way for more trustworthy, actionable AI in healthcare—bridging the critical gap between statistical performance and domain relevance.

ACKNOWLEDGMENT

This research has been funded by the German Federal Ministry of Education and Research (BMBF) under grant agreement no. 13FH5E11IA (CoHMed/NIO). Responsibility for the content of this publication lies with the author.

REFERENCES

[1] J. M. Duran and K. R. Jongsma, "Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical ai", *Journal of Medical Ethics*, vol. 47, no. 5, pp. 329–335, 2021. DOI: 10.1136/medethics-2020-106820.

[2] T. P. Quinn, S. Jacobs, M. Senadeera, V. Le, and S. Coghlan, "The three ghosts of medical ai: Can the black-box present deliver?", *Artificial Intelligence in Medicine*, vol. 124, p. 102 158, 2022, ISSN: 0933-3657. DOI: 10.1016/j.artmed.2021.102158.

[3] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions", in *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 2017.

[4] A. Bilal, D. Ebert, and B. Lin, "Llms for explainable ai: A comprehensive survey", *ACM Transactions on Intelligent Systems and Technology*, 2025, March 2025 edition.

[5] D. Martens, J. Hinns, C. Dams, M. Vergouwen, and T. Evgeniou, "Tell me a story! narrative-driven xai with large language models", *arXiv preprint*, 2023. eprint: 2309.17057.

[6] A. Zytek, S. Pido, S. Alnegheimish, L. Berti-Équille, and K. Veeramachaneni, "Explingo: Explaining ai predictions using large language models", in *IEEE Big Data Conference*, 2024. eprint: 2412.05145.

[7] J. Burton, N. A. Moubayed, and A. Enshaei, "Natural language explanations for machine-learning classification decisions", in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2023, pp. 1–9.

[8] M. A. Kadir, A. Mosavi, and D. Sonntag, "Evaluation metrics for xai: A review, taxonomy, and practical applications", in *27th IEEE International Conference on Intelligent Engineering Systems (INES)*, 2023, pp. 111–124. DOI: 10.1109/INES59282.2023.10297629.

[9] K. Matton, R. Ness, J. Guttag, and E. Kiciman, "Walk the talk? measuring the faithfulness of large language model explanations", in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[10] Y. Ming *et al.*, "Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows"", *arXiv preprint*, 2024. eprint: 2410.03727.

[11] N. Kroeger, D. Ley, S. Krishna, C. Agarwal, and H. Lakkaraju, "Are large language models post hoc explainers?", in *Robustness of Few-/Zero-Shot Learning Workshop @ NeurIPS 2023*, 2023.

[12] T. Lanham *et al.*, "Measuring faithfulness in chain-of-thought reasoning", *arXiv preprint*, 2023. eprint: 2307.13702.

[13] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: A method for automatic evaluation of machine translation", in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318. DOI: 10.3115/1073083.1073135.

[14] C. Lin, "Rouge: A package for automatic evaluation of summaries", in *Proceedings of the Workshop on Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81.

[15] L. Zhong, J. Wu, Q. Li, H. Peng, and X. Wu, "A comprehensive survey on automatic knowledge graph construction", *ACM Computing Surveys*, vol. 56, no. 4, 2024. DOI: 10.1145/3618295.

[16] F. Belleau, M. Nolin, N. Tourigny, A. Rigault, and J. Morissette, "Bio2rdf: Towards a mashup to build bioinformatics knowledge systems", *Journal of Biomedical Informatics*, vol. 41, no. 5, pp. 706–716, 2008. DOI: 10.1016/j.jbi.2008.03.004.

[17] Z. Chen, J. Chen, A. K. Singh, and M. Sra, "Xplainllm: A knowledge-augmented dataset for reliable grounded explanations in llms", in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore: Association for Computational Linguistics, 2024, pp. 7578–7596.

[18] Y. Gao *et al.*, "Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study", *JMIR AI*, vol. 4, e58670, 2025. DOI: 10.2196/58670.

[19] A. Khediri, H. Slimi, A. Yahiaoui, and M. Derdour, "Enhancing machine learning model interpretability in intrusion detection systems through shap explanations and llm-generated descriptions", in *6th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, 2024. DOI: 10.1109/PAIS62114.2024.10541168.

[20] L. P. Meyer *et al.*, "Llm-assisted knowledge graph engineering: Experiments with chatgpt", in *First Working Conference on Artificial Intelligence Development for a Resilient and Sustainable Tomorrow (AIDRST 2023)*, C. Zinke-Wehlmann and J. Friedrich, Eds., ser. Informatik aktuell, Wiesbaden, Germany: Springer Vieweg, 2024. DOI: 10.1007/978-3-658-43705-3_8.

[21] F. Gaber *et al.*, "Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis", *npj Digital Medicine*, vol. 8, p. 263, 2025. DOI: 10.1038/s41746-025-01684-1.

[22] Onkopedia Guidelines, "Colon carcinoma — onkopedia guideline", Accessed 2025-09, 2025, [Online]. Available: https://www.onkopedia.com/.

[23] M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie, "A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation", *Journal of Chronic Diseases*, vol. 40, no. 5, pp. 373–383, 1987. DOI: 10.1016/0021-9681(87)90171-8.

[24] R. G. Newcombe, "Two-sided confidence intervals for the single proportion: Comparison of seven methods", *Statistics in Medicine*, vol. 17, no. 8, pp. 857–872, 1998.

[25] M. Afshar, Y. Gao, D. Gupta, E. Croxford, and D. Demner-Fushman, "On the role of the umls in supporting diagnosis generation: Differential diagnoses proposed by large language models", *Journal of Biomedical Informatics*, vol. 157, p. 104 707, 2024. DOI: 10.1016/j.jbi.2024.104707.

Explaining the Medical Record: a Research Agenda for Non-medical Practitioners

Ray B. Jones, Aled Jones, Sally Abey, Patricia Schofield, Joanne Paton, Jill Shawe, Jenny Freeman, Avril Collinson, Nicholas Peres, John Downey, Sheena Asthana
University of Plymouth, PL4 8AA, United Kingdom
Emails: ray.jones@plymouth.ac.uk; aled.jones@plymouth,ac,uk; sally.abey@plymouth.ac.uk; patricia.schofield@plymouth.ac.uk; joanne.paton@plymouth.ac.uk; jill.shawe@plymouth.ac.uk; jenny.freeman@plymouth.ac.uk; avril.collinson@plymouth.ac.uk; nicholas.peres@plymouth.ac.uk; john.downey@plymouth.ac.uk; sheena.asthana@plymouth.ac.uk.

Abstract - This paper proposes a research agenda exploring how Generative Artificial Intelligence (GAI) can help explain patient medical records, particularly to the patients of non-medical practitioners. While patient access to records is expanding globally, little is known about how this access supports care beyond primary care doctors, or how GAI tools like ChatGPT may assist in interpretation. We outline key research questions and argue for co-designed solutions that include nurses, midwives, and allied health professionals to ensure accessible, equitable, and scalable approaches to explainability in digital health.

Keywords- explainability; patient access to records; research agenda; non-medical practitioners.

I. INTRODUCTION

Medical records were originally developed in the 18th and 19th centuries, primarily as an aide-mémoire for clinicians to support diagnosis, monitor treatment, and facilitate communication between healthcare professionals, not as documents intended for patients themselves. During the 1960s and 1970s researchers and practitioners began to suggest that patients could benefit from access to their records or hold shared care records [1], for example, in diabetes or hypertension [2]. As technology developed opportunities arose to share computer-produced summaries, for example, a clinical system for diabetes that produced records for hospital, GP and patient [3][4]. Use of this problem-oriented record showed that doctors were not always ready to share all problem-list entries with their patients [5][6]. On the other hand, in some situations such as antenatal care [7], clinicians were prepared to 'hand over' a complete paper medical record for women to look after.

In the 1990s we saw attempts to explain medical records to patients including the development of 'lay dictionaries' to 'translate' medical problems [8][9] as well as AI approaches to construct explanations [10][11] and showed that explanations based on their medical record were preferred to more generic information [12][13]. Randomised trials in the 1990s and 2000s [12][14][15] showed that giving patients access to their record with some type of explanation was of benefit. For example, a computer-produced paper record of the medical record with quality relevant information was more likely to be shared by cancer patients with their family than just the general information. This helped reduce patient anxiety [12].

More recently, a 2020 systematic review of patient access to medical records found that sharing electronic records with patients improved medicine safety and often reduced healthcare use, including fewer hospital visits and appointments [16]. However, an editorial by Sarkar et al [17] argued that the impact of patient access depends heavily on implementation. Contextual factors such as digital literacy, language, and clinical workflows must be considered, or else the benefits may be offset by increased clinician burden and exacerbated inequalities [18].

In section 2 we describe current practice, in section 3 the changing health information landscape in the UK, in section 4 we describe research questions about explaining medical records to patients, in section 5 we focus on under-researched areas and draw conclusion in section 6.

II. CURRENT PRACTICE

Progress in this area had been slow until recently, but patients in at least 30 countries now have some level of access to their records. Online routine access to medical records has demonstrated benefits including patient empowerment, reducing inefficiencies, error correction, and better shared decision making [19-21].

However, the degree of routine implementation differs. In the UK, patients were expected to gain prospective access to new data in their primary care records, including letters and consultations, from October 2023. However, a recent study [22] of 400 GPs in England revealed that in 2023 only 33% supported patient access to records. Most GPs felt that patients would worry more (91%) or find records confusing (85%). While many acknowledged potential patient benefits, most believed that online record access would increase their workload. Qualitative analysis [23] echoed these concerns among other primary care staff. Clinicians are concerned that patients will not understand their records.

III. THE CHANGING UK HEALTH INFORMATION LANDSCAPE

The NHS 10-Year Plan sets out a vision for a digitally enabled, personalised, and prevention-focused health service, emphasising the shift of care closer to home and the importance of empowering individuals to manage their own health. Achieving this vision requires not only giving patients access to their health records, but also ensuring they can understand and use that information effectively [24].

In the UK as elsewhere, the digital health landscape is evolving rapidly, both in terms of access to general health information and the development of personal health records. High-quality health information is widely available from trusted sources such as the NHS [26], Mayo Clinic [27], NICE [28], as well as peer-reviewed medical journals. This information is increasingly being accessed, summarised, and transformed by GAI tools such as ChatGPT.

Meanwhile, personal health records, created through interactions with frontline systems in general practice and community care (e.g., EMIS [29] and SystmOne [30]), as well as hospital systems (e.g., Cerner [31] and Epic [32]), are being extracted into patient-facing platforms such as the NHS App [33]. These records may also feed into shared care records for care planning and potential future patient access (e.g., via systems like Orion [34] and Black Pear [35]). Patients may therefore engage with digital health in different ways: using public websites or AI tools independently or verifying their clinical data through patient portals, then exploring it via GAI. Some health IT providers are beginning to integrate, or plan to integrate, GAI directly into their patient portal platforms. For example, Epic is working with Microsoft/OpenAI to embed GAI into clinician workflows and patient portals and NHS England is exploring how GAI might be used in the NHS App and other digital services.

GAI tools offer new opportunities to make medical records more accessible by translating clinical jargon into lay language, providing context-specific explanations, supporting conversational queries, and generating personalised summaries. These tools may enhance patient understanding, engagement, and self-management, especially when integrated with voice interfaces or patient portals. However, public-facing GAI tools also carry significant risks. They may generate incorrect or misleading information ("hallucinations"), lack source traceability, pose privacy concerns if sensitive data is shared outside secure systems, and exacerbate inequalities among patients with low digital literacy or poor internet access. Without safeguards and careful integration into clinical workflows, GAI may increase anxiety or misunderstandings rather than empowering patients. Research is therefore needed to explore how GAI can be safely and effectively deployed in real-world health contexts, particularly for non-medical practitioners and the populations they support.

IV. RESEARCH QUESTIONS ABOUT EXPLAINING MEDICAL RECORDS TO PATIENTS.

We could divide research questions about medical records into three categories:

- ‘Micro’ level, the explainability of the record, exploring which types of explanation are preferred or are more useful.
- ‘Meso’ level, whether patients want to use portals and whether their use and GAI affects the practitioner-patient relationship, and
- ‘Macro’ level, how this transformation can affect patient outcomes and possible changes to care processes, such as the shift from acute to community care and the focus towards health promotion and disease prevention [36].

Micro questions might include: How much do patients need their medical record if they know enough to ask a GAI for explanation? Will software developers build in GAI to their systems? Will this be more secure than patients using information from their online records to query a GAI? If NHS App builds in GAI will patients use that or still use independent GAI? What about the digitally disadvantaged? How should GAI adapt explanations to the knowledge level of the patient? Should the priority be on giving voice AI access to medical records so that those with no internet access or lack of skills can use the telephone to find out more?

At the ‘Meso’ level, questions are focussed on how we develop the triad of patient-practitioner and AI? What staff training is needed? How can practitioners collaborate with patients who turn up with lists or cite papers or GAI? How can practitioners support patients who do not use the Internet? How can practitioners assess their patients’ IT abilities and knowledge? How might this approach need to be adapted for some categories of patients such as the cognitively impaired? How do practitioners feel about patients reading and interpreting their notes—especially sensitive or nuanced ones (e.g. mental health, pain, uncertainty)? Does transparency change clinical documentation practices (e.g., tone, completeness, candour)? What are the risks and benefits of giving access to records in real time versus following clinician review or filtering? How do we introduce this topic to the curriculum of doctors, nurses, and other health professionals?

At the Macro level, NHS level questions are concerned with the most scalable and cost-effective methods for explaining records (e.g., automated summaries vs clinician review vs chatbot support)? How can health systems measure ‘understanding’ as an outcome of record-sharing interventions? Will these developments increase or decrease health inequalities?

V. UNDER RESEARCHED AREAS

In the English NHS, there are approximately 172,000 doctors (134,000 hospital doctors and 38,000 full-time equivalent GPs). However, there are some 372,000 nurses and midwives, and over 200,000 Allied Health Professionals (AHPs) (healthcare professionals other than doctors and nurses) from 14 professions (such as physiotherapy, podiatry, dietetics) working across community, primary, and secondary care. AHPs deliver over 208 million patient contacts annually [36]. Yet, most research into patient online access to their records has been in primary care and with GPs. Very little is known about nursing or AHPs’ or patients’ attitudes to patient access to their records or the use of GAI in non-medical clinical situations. For example, a recent scoping review of patient-accessible electronic health records [37] identified 66

studies, with none addressing nursing or AHP attitudes or GAI use in those settings.

We propose that the research questions outlined above regarding the most effective ways to explain medical records, could be more widely explored at micro, meso and macro levels, through co-design with patients and practitioners in non-medical disciplines. These include antenatal care, nurse-led pain clinics, physiotherapy, podiatry, and dietetics.

- Antenatal care has the longest history of providing patients with access to their records [7]. It continues to lead in shared record practices, with handheld notes and digital maternity apps now widely used.
- Pain clinics, particularly those led by nurses, are more cautious. While some services have begun to share care plans and symptom-tracking tools through patient portals, concerns remain about the risk of patients misinterpreting complex pharmacological or psychological data.
- Podiatry, especially within diabetes care, is seeing a growing use of digital platforms. These integrate podiatry notes into diabetes pathways and offer patients access to wound images, self-care advice, and foot health monitoring. However, access remains inconsistent.
- Dietetics is at a transitional stage. Patients are increasingly using digital tools to track dietary intake and receive tailored plans. There are also new digital platforms evolving such as MyRenalCare where clinicians including dietitians support the patient. Yet access to dietetic records is still limited, and documentation is not routinely shared or integrated across systems.
- Physiotherapy shows similar variability. Some integrated musculoskeletal pathways allow patients to access structured exercise plans and outcome data via apps like getUBetter or PhysiApp. However, routine access to clinical notes is uncommon, and many departments still rely on paper records or standalone systems.

Overall, progress toward shared records and digital self-management tools across these disciplines is uneven. There is a mix of promising developments and significant gaps. However, this inconsistency presents an opportunity: it offers researchers a diverse range of environments in which to explore and evaluate innovative approaches.

VI. CONCLUSION

Now is the time for a major change towards using AI to explain and interpret the content of a patient's medical record to the patient themselves. But we need (i) to switch attention to the under-researched areas of nursing and AHPs and (ii) to work with both practitioners and patients to co-design the convergence of patient access and GAI to empower patients to self-manage their condition and get what they need from their clinical consultation. Co-design is the only approach which identifies the needs and concerns of both groups (HCPs and patients) and enables them to work together in developing and sharing an optimum approach

We now need collaborative design between patients and practitioners to adapt these technologies effectively within clinical workflows. Without such work, we risk missing opportunities for improvement and compounding access disparities. This research proposes co-design approaches, including the development of solutions such as voiceAI telephone interfaces, to ensure these tools are usable, equitable, and aligned with NHS real-world needs.

Improvements in technology such as patient portals and GAI, may make it possible to improve patient autonomy, accelerate the switch from acute to community care, focus on health promotion and disease prevention. and reduce practitioner workloads. However, practitioners are concerned that the integration of AI and the potential need for deeper conversations with patients will add additional time pressures and create inefficiencies as conversations are misdirected to discuss strong preconceptions and conflicting advice, with some patient groups feeling empowered (but perhaps misinformed) while the more digitally excluded suffer even greater disadvantage.

To realise the benefits of patient access to records, particularly in community-based care, approaches must be co-designed by patients and practitioners and focus on inequalities. Despite extensive research in primary care and some in hospital settings, there has been virtually no exploration of patient access in collaboration with non-medical practitioners, apart from longstanding antenatal care research [7]. To unlock the full potential of patient-accessible records and generative AI, we must expand our research lens beyond doctors and engage the full breadth of the healthcare workforce and the patients they serve.

REFERENCES

[1] M. L. Gilhooly and S. M. McGhee, "Medical records: practicalities and principles of patient possession," Journal of Medical Ethics, vol. 17, no. 3, pp. 138–143, 1991.

[2] S. Ezedum and D. N. S. Kerr, "Collaborative care of hypertensives using a shared record," British Medical Journal, vol. 2, no. 6099, p. 1402, 1977.

[3] R. B. Jones, A. J. Hedley, and I. Peacock, "A patient held record and new methods of long term care for patients with diabetes mellitus," in 8th International Congress on Health Records, The Hague, Netherlands, pp 357-372, 1980.

[4] R. B. Jones et al., "A Computer-Assisted Register and Information-System for Diabetes," Methods of Information in Medicine, vol. 22, no. 1, pp. 4–14, 1983.

[5] R. B. Jones and A. J. Hedley, "Patient-Held Records - Censoring of Information by Doctors," Journal of the Royal College of Physicians of London, vol. 21, no. 1, pp. 35–38, 1987.

[6] R. B. Jones, A.J. Hedley, S.P. Allison, RB Tattersall, "Censoring of Patient-Held Records by Doctors," Journal of the Royal College of General Practitioners, vol. 38, no. 308, pp. 117–118, 1988.

[7] D. Elbourne, M.Richardson, I.Chalmers, L.Waterhouse, E.Holt "The Newbury Maternity care study – a randomized controlled trial to assess a policy of women holding their own obstetric records," British Journal of Obstetrics and Gynaecology, vol. 94, no. 7, pp. 612–619, 1987.

[8] R. Jones and P. Sandham, “A lay axis to the Read Codes?,” British Journal Healthcare Computing, pp. 30–31, 1994.
[9] S. M. McGhee, E. Symington, R. Jones, T. Hedley, G. McInnes, “Evaluation of a shared-care scheme for hypertension,” in Current Perspectives in Health Computing ’89: Conference & Exhibition, Harrogate, 12–14 April 1989, eds. Jane Duncan & Jacqueline Guinnane, published by British Journal of Healthcare Computing (BJHC Books). pp 19-21,1989.
[10] A. J. Cawsey, K. A. Binsted, and R. B. Jones, “An on-line explanation of the medical record to patients via an artificial intelligence approach,” in Current Perspectives in Healthcare Computing, Harrogate: BJHC Limited, 1995.
[11] K. Binsted, A. Cawsey, and R. Jones, “Generating Personalised Patient Information Using the Medical Record,” Lecture Notes in Artificial Intelligence, pp. 29–41, 1995.
[12] R. Jones et al., “Randomised trial of personalised computer based information for cancer patients,” British Medical Journal, vol. 319, no. 7219, pp. 1241–1247, 1999.
[13] R. Jones, J. Pearson, A. Cawsey, A. Barrett A, “Information for patients with cancer: does personalisation make a difference? Pilot study results and randomised trial in progress,” Journal of the American Medical Information Association (Symposium Supplement), pp. 423–427, 1996.
[14] R. B. Jones et al., “Randomised trial of personalised computer based information for patients with schizophrenia,” BMJ, vol. 322, no. 7290, pp. 835–840, 2001.
[15] R. B. Jones et al., “Effect of different forms of information produced for cancer patients on their use of the information, social support, and anxiety: randomised trial,” BMJ, vol. 332, no. 7547, pp. 942–946A, 2006.
[16] A. L. Neves, L.Freise, L.Laranjo, A.W. Carter, A. Darzi, E. Mayer, “Impact of providing patients access to electronic health records on quality and safety of care: a systematic review and meta-analysis,” BMJ Quality & Safety, vol. 29, no. 12, pp. 1019–1032, 2020.
[17] U. Sarkar and C. Lyles, “Devil in the details: understanding the effects of providing electronic health record access to patients and families,” BMJ Quality & Safety, vol. 29, no. 12, pp. 965–967, 2020.
[18] A. Turner et al., “Unintended consequences of patient online access to health records: a qualitative study in UK primary care,” British Journal of General Practice, vol. 73, no. 726, pp. E67–E74, 2023.
[19] C. Blease et al., “The benefits and harms of open notes in mental health: A Delphi survey of international experts,” PLoS One, vol. 16, no. 10, e0258056, 2021.
[20] C. Blease, I. G. Cohen, and S. Hoffman, “Sharing Clinical Notes Potential Medical-Legal Benefits and Risks,” JAMA, vol. 327, no. 8, pp. 717–718, 2022.
[21] A. Hannan and F. Webber, “Towards a Partnership of Trust,” in 4th Conference of the International-Council-on-Medical-and-Care-Compunetics, Amsterdam, Netherlands: IOS Press.pp 108-116.
[22] C. R. Blease et al., “Experiences and opinions of general practitioners with patient online record access: an online survey in England,” BMJ Open, vol. 14, no. 1, p. 10, 2024.
[23] C. Blease et al., “Patient Online Record Access in English Primary Care: Qualitative Survey Study of General Practitioners' Views,” Journal of Medical Internet Research, vol. 25, p. 13, 2023.
[24] S. Asthana, R. Jones, and R. Sheaff, “Why does the NHS struggle to adopt eHealth innovations? A review of macro, meso and micro factors,” BMC Health Services Research, vol. 19, no. 1, pp 1-7, 2019.
[25] Department of Health and Social Care (2023a) Fit for the future: 10-year health plan for England. Executive Summary. Available at: https://www.gov.uk/government/publications/10-year-health-plan-for-england-fit-for-the-future/fit-for-the-future-10-year-health-plan-for-england-executive-summary [Accessed 4 Aug. 2025].
[26] NHS https://www.nhs.uk/ (Accessed 28 August 2025)
[27] The Mayo Clinic https://www.mayoclinic.org/diseases-conditions (Accessed 28 August 2025)
[28] National Institute for Health and Social Care Excellence (NICE) https://www.nice.org.uk/what-nice-does/our-guidance/about-evidence-summaries (Accessed 28 August 2025)
[29] EMIS https://www.emishealth.com/emis-web (Accessed 28 August 2025)
[30] SystmOne https://tpp-uk.com/products/ (Accessed 28 August 2025)
[31] Cerner https://docs.oracle.com/en/cloud/paas/access-governance/ocoir/ (Accessed 28 August 2025)
[32] EPIC https://www.epic.com/ (Accessed 28 August 2025)
[33] NHSApp https://www.nhsapp.service.nhs.uk/login (Accessed 28 August 2025)
[34] Orion https://orionhealth.com/uk/ (Accessed 28 August 2025)
[35] Black Pear https://blackpear.com/ (Accessed 28 August 2025)
[36] NHS England, “The Allied Health Professions Strategy for England: AHPs Deliver (2022–2027),” 2022. [Online]. Available: https://www.england.nhs.uk/ahp/allied-health-professions-strategy-for-england/. [Accessed: 4-Aug-2025].
[37] T. Kariotis; M. Prictor, K. Gray, S. Chang, “Patient-Accessible Electronic Health Records and Information Practices in Mental Health Care Contexts: Scoping Review,” Journal of Medical Internet Research, vol. 27, p. 26, 2025.