P1 — Explainability in Ranking Algorithms

Explainable and Trustworthy AI Course

Politecnico di Torino - 2024/2025

Reference teachers: Eliana Pastor, Eleonora Poeta

Project. This project aims to analyze existing approaches to explainability in ranking algorithms and develop or enhance methods for generating human-understandable explanations for rankings to improve interpretability and user trust of ranked outputs across domains such as search engines, recommendation systems, and decision-support tools.

Overview.

Ranking algorithms [3] are central to numerous applications, including information retrieval, recommender systems, and hiring or admissions platforms. Despite their widespread adoption, the opaque nature of these models raises concerns about transparency, fairness [8], and user trust. Explainability in ranking aims to clarify why certain items are ranked higher than others, helping users make informed decisions and developers diagnose system behavior. Several approaches have emerged, proposing explanation techniques such as feature attribution [1, 2, 4], counterfactual explanations [6, 5, 7]. However, no distinct analysis emerges from this approach, and none addresses all aspects. This project addresses the need for a comprehensive analysis of explanation techniques tailored to the ranking context and for developing or enhancing an existing approach.

Goal.

The project aims to review existing methods for explainability in ranking algorithms systematically, identify current research gaps, and propose improved or novel methods for generating human-understandable explanations. It further seeks to evaluate the impact of these explanations on user trust, decision-making, and system adoption through empirical analysis.

Required analysis, implementation, and evaluation.

- **Literature Review.** Conduct a systematic review of existing explainability methods for ranking algorithms.

- **Identification of Research Gaps.** Identify key gaps in current methods, such as scalability, generalizability across domains, or alignment with human reasoning.
- **Implementation.** Develop and improve existing explainability techniques for rankings (e.g., via attribution, rule extraction, or contrastive explanations).
- **Evaluation.** Evaluate the effectiveness of the proposed methods through user studies or quantitative metrics.

References

[1] Alessandro Castelnovo et al. "Evaluative item-contrastive explanations in rankings". In: *Cognitive Computation* 16.6 (2024), pp. 3035–3050.

[2] Maria Heuss, Maarten de Rijke, and Avishek Anand. "RankingSHAP–Listwise Feature Attribution Explanations for Ranking Models". In: *arXiv preprint arXiv:2403.16085* (2024).

[3] Tie-Yan Liu et al. "Learning to rank for information retrieval". In: *Foundations and Trends® in Information Retrieval* 3.3 (2009), pp. 225–331.

[4] Venetia Pliatsika et al. *ShaRP: A Novel Feature Importance Framework for Ranking.* 2025. arXiv: `2401.16744 [cs.AI]`. URL: `https://arxiv.org/abs/2401.16744`.

[5] Joel Rorseth et al. "Credence: Counterfactual explanations for document ranking". In: *2023 IEEE 39th International Conference on Data Engineering (ICDE).* IEEE. 2023, pp. 3631–3634.

[6] Mozhgan Salimiparsa. "Counterfactual Explanations for Rankings." In: *Canadian AI.* 2023.

[7] Juntao Tan et al. "Counterfactual explainable recommendation". In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management.* 2021, pp. 1784–1793.

[8] Meike Zehlike, Ke Yang, and Julia Stoyanovich. "Fairness in ranking: A survey". In: *arXiv preprint arXiv:2103.14000* (2021).

P2 — Evaluating Explanations and their Sensitivity

Explainable and Trustworthy AI Course

Politecnico di Torino - 2024/2025

Reference teachers: Eliana Pastor, Eleonora Poeta

Project. This project aims to explore the evaluation of explanation methods, focusing on how explanation quality is affected by changes in data, model architecture, and training conditions, with the aim of assessing and improving the reliability and generalizability of explainability techniques.

Overview.

Explanation methods are designed to offer insights into machine learning model decisions. However, these methods often lack robustness and consistency when applied across different models or under minor input perturbations [1, 2, 3, 5, 4, 6, 8, 7, 9]. These variations raise concerns about whether current explainability tools truly reflect stable and generalizable properties of a model or its data. This project addresses the need for systematic approaches to evaluate explanation quality and explore whether explanations can be made more robust across models and settings.

Goal.

The aim of this project is to evaluate the behavior of explanation methods across diverse modeling conditions and to investigate strategies for improving the consistency and robustness of explanations. The project will consider a set of explanation methods and will examine their capacity to produce reliable insights across different scenarios. It aims to assess whether current explanation methods offer insights when subjected to varying model types, training settings, or slight modifications to the input data. In doing so, the project will also consider how evaluation strategies for explanations can be broadened to capture more general interpretability properties.

Required analysis, implementation, and evaluation.

- **Literature Review.** Conduct a review of existing evaluation metrics for explanations, focusing on criteria such as robustness, fidelity, stability, and generalizability. Examine also performed analysis evaluating such aspects.

- **Identification of Research Gaps.** Identify key limitations in how current explanation methods are evaluated and the extent to which these methods provide consistent insights across different training runs, models, or data configurations.

- **Implementation.** Apply selected explanation methods to a variety of model types and training configurations. Introduce experimental setups that test explanation variability under model perturbations, data changes, or re-training. Optionally, propose modifications to existing methods to enhance robustness or interpretability.

- **Evaluation.** Develop or adapt a general benchmarking strategy to assess explanation behavior across multiple conditions. Focus on characterizing explanation behavior, using or proposing metrics that reflect changes in explanations across models or inputs, and assessing to what extent explanations remain meaningful and stable under variation.

References

[1] Elvio Amparore, Alan Perotti, and Paolo Bajardi. "To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods". In: *PeerJ Computer Science* 7 (2021), e479.

[2] Francesco Bodria et al. "Benchmarking and survey of explanation methods for black box models". In: *Data Mining and Knowledge Discovery* 37.5 (2023), pp. 1719–1778.

[3] Oana-Maria Camburu et al. "Can I trust the explainer? Verifying post-hoc explanatory methods". In: *arXiv preprint arXiv:1910.02065* (2019).

[4] Sanjoy Dasgupta, Nave Frost, and Michal Moshkovitz. "Framework for evaluating faithfulness of local explanations". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 4794–4815.

[5] Ann-Kathrin Dombrowski et al. "Towards robust explanations for deep neural networks". In: *Pattern Recognition* 121 (2022), p. 108194.

[6] Cheng-Yu Hsieh et al. "Evaluations and methods for explanation through robustness analysis". In: *arXiv preprint arXiv:2006.00442* (2020).

[7] Giambattista Parascandolo et al. "Learning explanations that are hard to vary". In: *arXiv preprint arXiv:2009.00329* (2020).

[8] Dylan Slack et al. "Reliable post hoc explanations: Modeling uncertainty in explainability". In: *Advances in neural information processing systems* 34 (2021), pp. 9391–9404.

[9] Chih-Kuan Yeh et al. "On the (in) fidelity and sensitivity of explanations". In: *Advances in neural information processing systems* 32 (2019).

P3 - Automatic Subgroup Identification and Mitigation of Biases of ML Models

Explainable and Trustworthy AI Course

May 5, 2025

Reference teachers: Eleonora Poeta, Eliana Pastor

Project. This project aims to investigate and develop mitigation strategies to reduce performance disparities targeting underperforming subgroups discovered through automatic analysis.

Overview.

Machine learning models may exhibit disparities in performance across different population subgroups [10]. Identifying these underperforming subgroups is critical for improving model fairness and robustness. These disparities can stem from data imbalance, spurious correlations, or insufficient model capacity to generalize across diverse linguistic patterns. Prior work has identified underperforming subgroups using methods like clustering, frequent pattern mining, or metadata analysis [1, 3, 8, 9]. However, mitigation efforts frequently rely on predefined subgroups or simple debiasing techniques, which may not fully resolve deeper performance gaps [7, 14].

This project explores the development of effective mitigation strategies after the identification of problematic subgroups. Potential methods include data augmentation, loss reweighting, regularization techniques, and contrastive learning [2, 4, 5, 6, 12, 13]. Moreover, model-agnostic approaches like Shapley value [11] analysis and its approximations [8] can guide mitigation by highlighting key features correlated with underperformance, enabling targeted intervention either during model training or in post-processing.

Goal.

The primary goal is to develop and evaluate strategies for mitigating performance disparities. The project will start with the identification of underperforming subgroups and focus on improving fairness and robustness through subgroup-aware interventions.

Required analysis, implementation, and evaluation.

- **Literature Review.** Survey existing mitigation techniques for bias, with a focus on subgroup-specific fairness interventions.
- **Identification of Research Gaps.** Identify open challenges in current mitigation approaches for addressing subgroup-level disparities. For instance, many methods rely on static, user-defined subgroup labels or focus only on some specific type of data.
- **Implementation.** Develop a mitigation pipeline that addresses subgroup performance gaps. This may involve: (i) Using Shapley values or similar attribution techniques to identify key features associated with subgroup underperformance. (ii) Applying targeted mitigation techniques, such as generating counterfactual examples, augmenting low-performing group samples, or training with subgroup-aware loss functions.
- **Evaluation.** Evaluate the impact of the proposed mitigation strategy. Assess improvements using both performance metrics and fairness criteria (e.g., performance in data subgroups).

References

[1] Yeounoh Chung et al. "Slice finder: Automated data slicing for model validation". In: *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE. 2019, pp. 1550–1553.

[2] Pranav Dheram et al. "Toward Fairness in Speech Recognition: Discovery and mitigation of performance disparities". In: *Proc. Interspeech 2022*. 2022, pp. 1268–1272. DOI: 10.21437/Interspeech.2022-10816.

[3] Alkis Koudounas et al. "Exploring subgroup performance in end-to-end speech models". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[4] Alkis Koudounas et al. "Mitigating Subgroup Disparities in Speech Models: A Divergence-Aware Dual Strategy". In: *IEEE Transactions on Audio, Speech and Language Processing* 33 (2025), pp. 883–895. DOI: 10.1109/TASLPRO.2025.3539429.

[5] Alkis Koudounas et al. "Prioritizing Data Acquisition For End-to-End Speech Model Improvement". In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2024, pp. 1–5. DOI: 10.1109/ICASSP48485.2024.10446326.

[6] Pranay K Lohia et al. "Bias mitigation post-processing for individual and group fairness". In: *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE. 2019, pp. 2847–2851.

[7] Oliver Niebuhr and Alexis Michaud. “Speech data acquisition: the underestimated challenge”. In: *KALIPHO-Kieler Arbeiten zur Linguistik und Phonetik* 3 (2015), pp. 1–42.

[8] Eliana Pastor, Luca De Alfaro, and Elena Baralis. “Looking for trouble: Analyzing classifier behavior via pattern divergence”. In: *Proceedings of the 2021 International Conference on Management of Data*. 2021, pp. 1400–1412.

[9] Svetlana Sagadeeva and Matthias Boehm. “Sliceline: Fast, linear-algebra-based slice finding for ml model debugging”. In: *Proceedings of the 2021 International Conference on Management of Data*. 2021, pp. 2290–2299.

[10] Nima Shahbazi et al. “Representation bias in data: a survey on identification and resolution techniques”. In: *ACM Computing Surveys* 55.13s (2023), pp. 1–39.

[11] Lloyd S Shapley et al. “A value for n-person games”. In: (1953).

[12] Irina-Elena Veliche and Pascale Fung. “Improving Fairness and Robustness in End-to-End Speech Recognition Through Unsupervised Clustering”. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[13] Zeyu Wang et al. “Towards fairness in visual recognition: Effective strategies for bias mitigation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 8919–8928.

[14] Yuanyuan Zhang et al. “Mitigating bias against non-native accents”. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 2022. 2022, pp. 3168–3172.

P4 — Explainability meets Adversarial Attacks: Leveraging Explanations to Create Intrinsically Robust Models

Explainable and Trustworthy AI Course

Politecnico di Torino - 2024/2025

Reference teachers: Gabriele Ciravegna, Eleonora Poeta

1 Project Overview

Adversarial attacks pose a significant threat to machine learning models, manipulating inputs to mislead predictions and compromising system reliability. Explainable AI (XAI) offers a powerful avenue to enhance model transparency, enabling the identification and mitigation of adversarial behaviors. This project explores XAI techniques to detect, understand, and counteract adversarial attacks within the domain of computer engineering, ensuring the robustness of AI-driven solutions.

Several works can be identified as foundational milestones for adversarial attack detection and defense mechanisms. Biggio et al. [2013], Szegedy et al. [2013] and Goodfellow et al. [2014] first introduced the idea of *Adversarial Examples* in machine learning, highlighting vulnerabilities of Support Vector Machines and Deep Learning models. Papernot et al. [2016] proposed one of the first and most effective defense methods, based on the concept of distillation as a learning technique to mitigate adversarial attacks. Madry et al. [2017] proposed to train models on adversarial samples to improve their generalization and intrinsic robustness. Xu et al. [2017] proposed feature squeezing to compare a DNN model's prediction on the original input with that on squeezed inputs to detect adversarial attacks.

Recent research has explored XAI-driven adversarial defense strategies Liu et al. [2021]. For instance, Fidel et al. [2020] explored the use of Shapley Additive Explanations (SHAP) values computed for the internal layers of a DNN classifier to discriminate between normal and adversarial inputs. Zhang et al. [2018], instead, proposed detecting adversarial perturbations directly starting from saliency maps. Using Concept-based XAI methods, Ciravegna et al. [2023] have shown that the violation of the logic rules explaining a model behavior can be used as a detection method of adversarial samples.

Starting from these works, the student involved in this project will aim to integrate further explore and integrate XAI methodologies for adversarial attack mitigation, fostering more secure and transparent AI.

2 Goal

The primary goal of this project is for students to acquire and demonstrate their theoretical knowledge and practical skills in XAI techniques and adversarial attack, their detection and mitigation. Students will analyze various adversarial attack strategies and their impact on AI models, investigate XAI methodologies to interpret and diagnose adversarial manipulations, develop countermeasures leveraging XAI-driven explanations for enhanced AI security, and evaluate the effectiveness of mitigation strategies through empirical analysis and experimentation.

3 Required Steps

3.1 Literature Review

Students will be required to conduct a thorough review of contemporary research on adversarial attacks in AI models, existing XAI approaches and assessing prior work integrating XAI with adversarial defense.

3.2 Identification of Research Gaps

The project requires students to analyze current adversarial mitigation strategies, identifying their limitations. Possibly, they might find underrepresented XAI techniques in adversarial contexts and formulate proposals that might fill these key gaps in the literature.

3.3 Implementation

The project will involve the design and development of AI models with integrated XAI components. It will also require the students to implement different types of adversarial attacks to evaluate the model robustness.

3.4 Evaluation

The effectiveness of the XAI defense will be measured by analyzing model performance before and after its application. Adversarial robustness metrics (such as the Robust Accuracy or the detection percentage) will be examined to assess their role in security improvements. Possibly, the results should be compared against at least an existing state-of-the-art adversarial defense mechanism.

References

B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *Machine learning and knowledge discovery in databases: European conference, ECML pKDD 2013, prague, czech Republic, September 23-27, 2013, proceedings, part III 13*, pages 387–402. Springer, 2013.

G. Ciravegna, P. Barbiero, F. Giannini, M. Gori, P. Lió, M. Maggini, and S. Melacci. Logic explained networks. *Artificial Intelligence*, 314:103822, 2023.

G. Fidel, R. Bitton, and A. Shabtai. When explainability meets adversarial learning: Detecting adversarial examples using shap signatures. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

N. Liu, M. Du, R. Guo, H. Liu, and X. Hu. Adversarial attacks and defenses: An interpretation perspective. *ACM SIGKDD Explorations Newsletter*, 23 (1):86–99, 2021.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

W. Xu, D. Evans, and Y. Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

C. Zhang, Z. Yang, and Z. Ye. Detecting adversarial perturbations with salieny. In *Proceedings of the 6th International Conference on Information Technology: IoT and Smart City*, pages 25–30, 2018.

P5 — Explainable-by-design Models for Autonomous Driving

Explainable and Trustworthy AI Course

Politecnico di Torino - 2024/2025

Reference teachers: Gabriele Ciravegna, Eleonora Poeta

1 Project Overview

Students engaging in this project will investigate explainable-by-design models for autonomous driving in order to enhance the transparency and trustworthiness of AI-driven decision-making systems. Autonomous vehicles often depend on deep learning for real-time decisions; however, the inherent opacity of these systems can raise critical safety, ethical, and regulatory concerns [Zablocki et al., 2022, Atakishiyev et al., 2024]. Through this project, students will develop models intrinsically equipped with explainable modules, thereby providing a clear rationale for autonomous actions.

Students will be required to implement at least one explainable approach—such as Concept Bottleneck Models, Prototype-Based Learning, or Attention-Based Interpretability—to elucidate key driving decisions [Lai et al., 2024]. For example, concept bottleneck models may enable the system to predict intermediate, human-defined concepts (e.g., "pedestrian crossing" or "vehicle stopping") before determining the action of the autonomous vehicle. Prototype-based approaches, in contrast, classify inputs by comparing them with representative reference examples; whereas attention-based models offer a visual breakdown of the most salient part of the input influencing the final classification.

For hands-on experimentation, students will utilize at least one large-scale autonomous driving dataset. Recommended datasets include the ROAD dataset [Singh et al., 2022], nuScenes [Caesar et al., 2020], the Waymo Open Dataset [Sun et al., 2020], and BDD100K [Yu et al., 2020]. These datasets provide a rich collection of frames (images) extracted from divers driving videos and labelled with the corresponding driving action taken by the driver.

2 Goal

Students will collaborate in groups to study, design and implement explainable-by-design autonomous driving models. They will be required to thoroughly present their results through a structured report. They will study and understand how explainable-by-design models can be applied in practical and sensitive scenarios. Also, they will gain hands-on experience with Explainable AI (XAI) techniques, selecting and implementing explainable-by-design models on large-scale autonomous-driving datasets.

3 Required Steps

3.1 Literature Review

Students will first perform a comprehensive review of Explainable AI (XAI) methodologies, particularly focusing explainable-by-design models and their application into practical scenarios such as the autonomous-driving one.

3.2 Identification of Research Gaps

Students are required to critically assess existing autonomous driving frameworks, identifying their limitations. Emphasis should be placed on the challenges of integrating real-time interpretability—particularly how current explainable-by-design models may be defined at the frame-level to also capture the temporal (video) context [Zhang et al., 2025]. Based on this analysis, students should formulate precise research questions and solutions to address the identified gaps.

3.3 Implementation

Students will implement at least one explainable model within a simulated autonomous driving pipeline. They may choose, for instance, a Concept Bottleneck Model, a Prototype-Based model, or an Attention-Based approach. In the case of Concept Bottleneck Models, students are encouraged to employ a pre-trained object detection model—such as the 3D-RetinaNet[1]—to predict objects in a given frame. This pre-trained model can serve as an effective means to furnish intermediate, human-interpretable concepts before the final decision is made.

Moreover, students will apply their chosen model to at least one dataset from the recommended list (ROAD, nuScenes, Waymo Open Dataset, BDD100K). This will allow for empirical validations under a variety of driving scenarios and environmental conditions. *Note: Students must ensure that the dataset they select is compatible with the chosen explainable-by-design model.*

[1] `https://github.com/gurkirt/3D-RetinaNet`

3.4 Evaluation

Students must rigorously evaluate the performance and interpretability of their models. This will involve comparing their approach against traditional black-box models using metrics that assess model performance, but also highlighting the advantage of explainable-by-design models.

References

S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *IEEE Access*, 2024.

H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

S. Lai, T. Xue, H. Xiao, L. Hu, J. Wu, N. Feng, R. Guan, H. Liao, Z. Li, and Y. Yue. Drive: Dependable robust interpretable visionary ensemble framework in autonomous driving. *arXiv preprint arXiv:2409.10330*, 2024.

G. Singh, S. Akrigg, M. Di Maio, V. Fontana, R. J. Alitappeh, S. Khan, S. Saha, K. Jeddisaravi, F. Yousefi, J. Culley, et al. Road: The road event awareness dataset for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1036–1054, 2022.

P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord. Explainability of deep vision-based autonomous driving systems: Review and challenges. *International Journal of Computer Vision*, 130(10):2425–2452, 2022.

B. Zhang, N. Song, X. Jin, and L. Zhang. Bridging past and future: End-to-end autonomous driving with historical prediction and planning. *arXiv preprint arXiv:2503.14182*, 2025.

P6 — Robustness in Medical Imaging Models: Towards Trustworthy AI Diagnostics

Explainable and Trustworthy AI Course

Politecnico di Torino - 2024/2025

Reference teachers: Eleonora Poeta, Gabriele Ciravegna

Project. This project focuses on exploring the robustness of AI models against Out-of-Distribution (OOD) inputs in Medical Imaging. It involves developing detection techniques to identify and mitigate these problems, thereby enhancing the robustness and trustworthiness of AI models used in medical diagnostics. The project emphasizes practical implementation and analysis on real-world medical imaging datasets (e.g., MRIs or CTs).

Overview.

In real-world clinical environments, AI models are frequently exposed to data that deviates from their original training distributions due to numerous sources of variability. These Out-of-Distribution (OOD) inputs can significantly degrade model performance and lead to unreliable predictions [1].

In the medical domain [3], the deployment of AI models that are not robust to such variations—whether demographic (e.g., age, gender, ethnicity) or technical (e.g., MRI scanners with differing field strengths or X-ray machines with varying resolutions)—is highly problematic due to the high-stakes nature of clinical decision-making [2, 4]. Models that fail to generalize across these variations risk producing biased or unsafe outputs, thereby compromising both diagnostic accuracy and patient safety.

Goal.

The main objectives are:

- Understanding challenges posed by OOD inputs in medical imaging applications.
- Investigating and implementing state-of-the-art OOD detection methods tailored for medical imaging tasks.
- Apply some mitigation strategy or evaluation of model robustness in a real-world scenario.

Required analysis, implementation, and evaluation.

- **Literature Review**. Conduct a systematic review of existing OOD detection methods, Robustness in Trustworthy AI, and application in the medical domain.
- **Identification of Research Gaps**. Highlight critical limitations in existing approaches, including the reliance on re-training to assess OOD behavior when models are exposed to data from varying institutions, imaging equipment, or patient populations.
- **Implementation**. Develop and improve existing techniques to evaluate or ensure the model's robustness.
- **Evaluation**.Evaluate the effectiveness of the proposed methods through user studies or quantitative metrics.

References

[1] Houssem Ben Braiek and Foutse Khomh. "Machine learning robustness: A primer". In: *Trustworthy AI in Medical Imaging*. Elsevier, 2025, pp. 37–71.

[2] Moritz Fuchs et al. "Navigating the unknown: out-of-distribution detection for medical imaging". In: *Trustworthy AI in Medical Imaging*. Elsevier, 2025, pp. 73–99.

[3] Zesheng Hong et al. "Out-of-distribution Detection in Medical Image Analysis: A Survey". In: *arXiv preprint arXiv:2404.18279* (2024). URL: `https://arxiv.org/abs/2404.18279`.

[4] Jonas Richiardi et al. "Domain shift, domain adaptation, and generalization: A focus on MRI". In: *Trustworthy AI in Medical Imaging*. Elsevier, 2025, pp. 127–151.