Addressing Explainability, Transparency and Interpretability Requirements of AI Models for ECG Analysis

Durande B K Nguifo[1], Gabriel T Zago[2], Rafael Pereira[2], Rodrigo V Andreão[2], Anthony Fleury[1]

[1] IMT Nord Europe, CERI SN, Institut Mines-Télécom, Lille, France
[2] Department of Electrical Engineering, Federal Institute of Espirito Santo, Vitória, Brazil

Abstract

Increased application of artificial intelligence (AI), particularly deep neural networks, in electrocardiogram (ECG) interpretation has highlighted critical concerns about the opacity of algorithmic decisions. The demand for explainability, interpretability, and transparency in medical AI has driven the development of explainable AI (XAI) techniques. This review aims to provide a comprehensive evaluation of the most relevant and promising XAI methods for ECG analysis, highlighting their methodological diversity and identifying the main challenges that must be addressed for clinical translation. Despite rapid advancements in XAI for ECG interpretation, significant gaps remain in clinical validation, personalization, and ethical governance. The findings underscore the need to integrate explainability with robust performance, transparency, and clinician-in-the-loop approaches. This review serves as a reference for future work that aims to bridge the gap between experimental research and clinically viable AI systems, ultimately promoting safer and more accountable deployment of AI in cardiology.

1. Introduction

Cardiovascular diseases (CVDs) are the main cause of death globally, accounting for approximately 17.9 million deaths each year. CVDs are a group of diseases of the heart and blood vessels and include heart diseases, cerebrovascular diseases, rheumatic heart diseases and other conditions. More than four out of five CVD deaths are due to heart attacks and strokes, and one third of these deaths occur prematurely in people under 70 years of age [1]. The diagnosis of the group of heart diseases is commonly based on the electrocardiogram (ECG), which is a simple, cheap, and noninvasive exam.

Computer systems based on artificial intelligence (AI) methods have been widely proposed to assist healthcare professionals in analyzing the ECG, offering an automatic and accurate interpretation of the exam and therefore the cardiac symptoms of the patient. Moreover, computer systems allow to detect prematurely abnormal conditions which in turn open the possibility to work on preventive care [2, 3].

In addition to the advantages of computer systems, the extensive use of AI methods for ECG analysis, particularly based on deep neural networks, has been followed by a great deal of concern about the ability to explain the decision made by clinicians [4]. The works in general prioritize performance. This issue has raised an increased attention towards explainable AI models (XAI), which are expected to address requirements of explainability, interpretability, and transparency to favor their adoption in clinical settings [5–8].

This paper presents a review of the most promising XAI approaches for ECG analysis and points out the main challenges. For this review, we selected 60 studies from the major databases. These articles were analyzed considering the type of model used (see Figure 1). We discuss the limitations of current approaches and the perspectives, aiming at pushing forward the interest and the adoption of XAI models for ECG analysis.

2. Materials and Methods

This section outlines the systematic methodology used for the literature review of XAI models applied to the analysis of ECG signals, specifically for time series classification. Our approach aimed to identify, analyze, and synthesize the most promising works in this field, highlighting the requirements for explainability, interpretability, and transparency to favor their clinical adoption.

2.1. Search Strategy and Data Sources

A systematic literature search was conducted across several major scientific databases to ensure complete coverage of the relevant literature. The databases consulted include: IEEEXplore, ScienceDirect, Scopus and Google Scholar. The search terms (keywords) used, alone or in combination, included: “XAI ECG”, “explainable AI

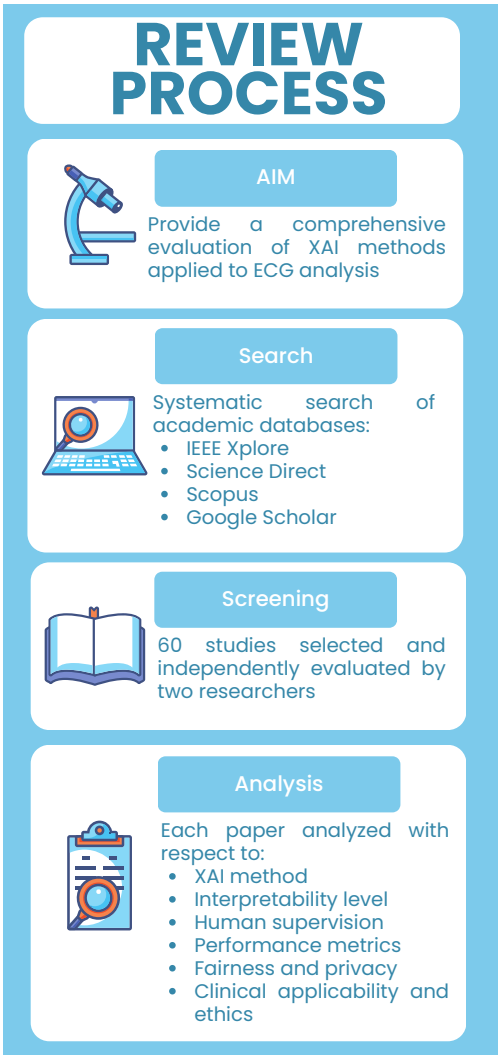


Figure 1. An overview of the review process

electrocardiogram", "interpretability cardiac arrhythmia", "transparency deep learning ECG", "explainable time series classification ECG". These terms have been adapted for each database to optimize results. The search focused specifically on works combining the concepts of XAI and ECG signal processing.

2.2. Study Selection Criteria

Identified articles were subject to a rigorous selection process based on the following inclusion and exclusion criteria:

Inclusion Criteria: (i) Research articles published in peer-reviewed journals or conference proceedings; (ii) Works presenting AI/machine learning models applied to the analysis and/or classification of ECG signals; (iii) Articles explicitly addressing the concepts of explainability, interpretability, and/or transparency of AI models for ECG analysis; (iv) Original studies describing methods, architectures, or performance evaluations of XAI systems for ECG data.

Exclusion Criteria: (i) Articles not directly related to ECG signal analysis (e.g., other physiological signals without ECG relevance); (ii) Articles not explicitly discussing AI explainability, interpretability, or transparency; (iii) Articles in languages other than English.

Study selection was carried out in two phases: first, a review of titles and abstracts to exclude irrelevant articles; second, a full-text review to confirm compliance with the inclusion criteria and analyze deeply the study performed.

2.3. Data Extraction and Analysis

For each selected article, relevant information has been systematically extracted using a predefined data collection form. This form allowed for the detailed collection of information on the methodological aspects and conclusions of each study, including:

- **Paper ID:** Article title and references.
- **Expert-Model Agreement:** Analysis of the concordance between the automatic method and the diagnosis of a medical expert.
- **Expert Adjustment:** Possibility for an expert to adjust the model decision after interpretation.
- **Decision Type:** Nature of the model decision (binary, fuzzy, probabilistic).
- **AI/ML Methods Employed:** Specific architectures and algorithms (e.g., CNN, RNN, LSTM, U-Net, Random Forest).
- **Performance Metrics:** Indicators used to evaluate the models (e.g., AUC, Precision, Sensitivity, Specificity, F-score, Accuracy).
- **XAI Performance:** Comparison of the performance of explainable methods versus traditional non-explainable methods, when applicable.
- **ECG Anomaly Classes:** Types of arrhythmia or ECG signal anomalies considered.
- **Databases:** Data sources used (e.g., MIT-BIH Arrhythmia Database, LUDB).
- **Patient Data Adaptation:** Ability of the model to adapt to individual patient-specific data.
- **Mechanisms of Explainability/Transparency:** How the model makes its decisions understandable (e.g., identification of specific signal regions, visualization of internal network operations).
- **Data Security/Privacy:** Implementation of data protection methods (e.g., anonymization, encryption, federated models).
- **Testing across Populations/Ethnicities:** Evaluation of model capacity to generalize.

• **Application Domain:** Model target (e.g., patient triage, queue reduction, remote care).
• **Error Responsibility:** Discussion of responsibility in case of diagnostic error.
• **Algorithm-Physician Discussion:** Possibility of case discussion between the algorithm's result and the clinician.

2.4. Common XAI Methods for Time Series Classification (ECG)

The increasing adoption of deep neural network models for ECG analysis has introduced a growing demand for XAI to ensure trustworthy of their decisions, particularly for time-series classification tasks like arrhythmia detection. This subsection details key XAI methods frequently employed or adapted for interpreting predictions on time-series data, specifically for ECG signals.

• **Local Interpretable Model-agnostic Explanations (LIME) [9]:** is a model-agnostic technique that provides local explanations by approximating the black-box model behavior around a specific prediction with a simpler and interpretable model. For time series, this involves creating disturbed versions of the input ECG signal (e.g., by masking or adding noise to specific segments or features over time) and observing how the model prediction changes. The resulting interpretable model then highlights which temporal segments or signal characteristics are most influential for that particular ECG classification.

• **Gradient-weighted Class Activation Mapping (Grad-CAM) [10] for Time Series:** Originally developed for convolutional neural networks (CNNs) in image processing, Grad-CAM can be effectively adapted for time series by applying it to 1D convolutional layers common in ECG analysis models. It generates a "saliency map" by using the gradients of the target class prediction with respect to the feature maps of a specific layer. For ECG time series, this map visually highlights the most important time points or segments within the waveform that drove the model's decision for a given diagnosis, such as identifying the specific peaks or intervals crucial for detecting an arrhythmia.

• **SHapley Additive exPlanations (SHAP)[11]:** is a unified framework based on cooperative game theory that assigns an "importance value" (Shapley value) to each feature for a given prediction. When applied to time series, SHAP quantifies the contribution of sample points or segments of an ECG signal to the output. This allows for a granular understanding of how specific temporal patterns (e.g., an abnormal P-wave duration or QRS morphology) positively or negatively influence the ECG classification.

• **Agnostic Local Explanation for Time Series Classification (LEFTIST) [12]:** adapts the framework LIME for time series classification. This method proposes the use of simplified "shapelets" as interpretable components, where these shapelets have a prefixed length and size across the time series. LEFTIST then provides a measure of feature importance for each of these identified shapelets, thereby explaining the model decision in the context of specific, salient time series patterns.

• **Counterfactual Explanations for Machine Learning on Multivariate Time Series Data (COMTE) [13]:** creates counterfactuals in multivariate time series by disturbing features of a time series with a heuristic algorithm. It adapts Wachter et al.'s [14] original formulation by replacing the point-wise distance function with a feature-based function. It is then possible to generate minimally different hypothetical time series that lead to an alternative classification, offering actionable insights into model decisions for complex multivariate time series.

• **Evolutionary Counterfactual Explanations for Time Series [15]:** adapts the concept of counterfactual explanations for both univariate and multivariate time series classification. It achieves this by formulating and optimizing a problem, and output distance. The optimization problem is then solved through the application of an evolutionary algorithm, specifically Non-dominated Sorting Genetic Algorithm (NSGA-II), which utilizes mutation operators that have been specifically adapted for the time series domain. The next section will discuss the results of our analysis.

3. Results

This comprehensive analysis of XAI methods in ECG reveals significant insights and gap between current research and clinical implementation requirements. Of 60 studies, only 38 were selected with all reviewer agreeing that XAI methods were used.

Clinical Integration Challenges: the most critical finding is the minimal integration between automated methods and clinical expertise. Only 9 out of 38 studies analyzed agreement with medical experts, while none provided mechanisms for expert adjustment of model outputs. This disconnect represents a fundamental barrier to real clinical adoption, as successful machine learning implementation requires physician-algorithm collaboration rather than replacement.

Methodological limitations: the predominance of binary decision-making approaches (13 studies) over probabilistic systems reflects an oversimplification of the diagnosis. Real-world scenarios require uncertainty quantification and graded assessments, particularly for borderline cases. Lack of fuzzy logic implementations also suggests a huge limitation in the decision paradigms.

Critical Gaps in Clinical Translation: several findings highlight significant barriers to real-world implementation:

Several critical gaps hinder real-world implementation:

Personalization: Near-complete absence of patient-specific adaptation (35 studies without) limits clinical accuracy across diverse physiological variations.

Security and Diversity: The lack of data security and privacy implementation (24 studies without any) is particularly concerning given the sensitive nature of medical data and increasing regulatory requirements. Only 2 studies implemented federated learning approaches, suggesting insufficient attention to privacy-preserving AI development, and the absence of multi-ethnic testing (28 studies) raises serious concerns about algorithmic bias and capacity of generalization across a diverse patient population.

Accountability and Responsibility We observe a lack of discussion upon the responsibility for errors (37 studies not addressing this) which represents a significant ethical and legal issue. Clear accountability frameworks are essential for AI deployment in real settings, particularly in high-stakes medical decision-making scenarios.

Future Directions The predominant reliance on CNN approaches (21 studies) and simple explainability mechanisms (ECG location identification: 11 studies) suggests technological tunnel vision. Integration of robust physician-AI frameworks, standardized evaluation protocols, and comprehensive bias mitigation strategies is essential. The journey toward clinically viable XAI requires interdisciplinary collaboration addressing technical performance, clinical utility, ethical governance, and regulatory compliance to bridge the gap between experimental research and safe, accountable cardiac care deployment.

4. Conclusion

In this paper, we presented a review of studies on XAI methods applied to ECG analysis. After paper selection, our study shows that, to date, it will be difficult to have automatic cardiac signal analysis by AI in real hospital settings. Indeed, our review of the state of the art highlighted that current studies fails on several key points, such as explainability of the automatic decision, lack of feedback of the healthcare professionals on the model's evolution, lack of security discussion, lack of diversity in the datasets, and lack of discussion on the responsibility considering the decision. All these points are mandatory to design real applications of these works in clinical settings.

Acknowledgements

This work was supported in part by grants EUVETUS from CAPES STIC AmSud (23-STIC-09) and CNPq.

References

[1] World Health Organization. Cardiovascular diseases. https://www.who.int/health-topics/cardiovascular-diseases/. On: 2025-08-17.

[2] Corral-Acero J, Margara F, Marciniak M, Rodero C, Loncaric F, Feng Y, Gilbert A, Fernandes JF, Bukhari HA, Wajdan A, et al. The 'digital twin'to enable the vision of precision cardiology. European Heart Journal 2020; 41(48):4556–4564.

[3] Gillette K, Gsell MA, Prassl AJ, Karabelas E, Reiter U, Reiter G, Grandits T, Payer C, Štern D, Urschler M, et al. A framework for the generation of digital twins of cardiac electrophysiology from clinical 12-leads ecgs. Medical Image Analysis 2021;71:102080.

[4] Al-Zaiti SS, Bond RR. Explainable-by-design: Challenges, pitfalls, and opportunities for the clinical adoption of ai-enabled ecg. Journal of Electrocardiology 2023;81:292–294.

[5] Li B, Qi P, Liu B, Di S, Liu J, Pei J, Yi J, Zhou B. Trustworthy ai: From principles to practices. ACM Computing Surveys 2023;55(9):1–46.

[6] Ayano YM, Schwenker F, Dufera BD, Debelee TG. Interpretable machine learning techniques in ecg-based heart disease classification: a systematic review. Diagnostics 2022;13(1):111.

[7] Neves I, Folgado D, Santos S, Barandas M, Campagner A, Ronzio L, Cabitza F, Gamboa H. Interpretable heartbeat classification using local model-agnostic explanations on ecgs. Computers in Biology and Medicine 2021; 133:104393.

[8] Bodini M, Rivolta MW, Sassi R. Opening the black box: interpretability of machine learning algorithms in electrocardiography. Philosophical Transactions of the Royal Society A 2021;379(2212):20200253.

[9] Ribeiro MT, Singh S, Guestrin C. " 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. 2016; 1135–1144.

[10] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision October 2019;128(2):336–359. ISSN 1573-1405.

[11] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 2017;30.

[12] Guillemé M, Masson V, Rozé L, Termier A. Agnostic local explanation for time series classification. In 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2019; 432–439.

[13] Ates E, Aksar B, Leung VJ, Coskun AK. Counterfactual explanations for multivariate time series. In 2021 International Conference on Applied Artificial Intelligence (ICAPAI). IEEE, 2021; 1–8.

[14] Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv JL Tech 2017;31:841.

[15] Höllig J, Kulbach C, Thoma S. Tsevo: Evolutionary counterfactual explanations for time series classification. In 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2022; 29–36.

Address for correspondence: Gabriel Tozatto Zago. Av. dos Sabiás, 330 - Morada de Laranjeiras, Serra - ES, 29166-630. gabriel.zago@ifes.edu.br