International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Hybrid BiLSTM-CNN Model with Attention and XAI (SHAP & LIME) for ECG Arrhythmia Classification Using PTB-XL

Maryam Shadan[1], Adicherla Sai Chaithanya[2], Chikoti Vaishali[3], Syed Mohammed Musharraf[4], Mujtaba Gulam Muqeeth[5]

Methodist College of Engineering and Technology, Student, Department of Computer Engineering, Abids, Hyderabad, Telangana, 500001, India
Email: *maryamshadan76[at]gmail.com*

Methodist College of Engineering and Technology, Student, Department of Computer Engineering, Abids, Hyderabad, Telangana, 500001, India
Email: *adicherlachaithu[at]gmail.com*

Methodist College of Engineering and Technology, Student, Department of Computer Engineering, Abids, Hyderabad, Telangana, 500001, India
Email: *chikotivaishali[at]gmail.com*

Methodist College of Engineering and Technology, Student, Department of Computer Engineering, Abids, Hyderabad, Telangana, 500001, India
Email*: syedmusharraf042[at]gmail.com*

Methodist College of Engineering and Technology, Associate Professor, Department of Computer Engineering, Abids, Hyderabad, Telangana, 500001, India
Email: *g.mujtaba[at]methodist.edu.in*

Abstract: ***Analysis of electrocardiograms (ECG) is vital to identify cardiological problems; however, traditional interpretation requires a lot of clinical knowledge and is influenced by inter-observer variability. In order to solve these problems, this research presents an interpretable deep learning framework that fuses Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM) networks, and an Attention mechanism for multi-label ECG classification. In addition to employing the PTB-XL dataset, the system commits to clinical trust by integrating explainability methods SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) to enhance transparency. The model reaches a macro F1-score of 0.369 and an ROC-AUC of 0.919, reflecting its high capability of simultaneous multi-class identification of different cardiac diseases. The SHAP–Attention superposition helps in providing easy-to-understand explanations, which enable the doctor to see the time-related and shape-related factors of the ECG waveform. This paper focuses on the importance of combining accuracy with explanation in AI healthcare systems and provides a robust, transparent method for computer-assisted cardiac diagnosis.***

Keywords: ECG Classification, SHAP, LIME, CNN-BiLSTM, Attention Mechanism, Deep Learning, PTB-XL Dataset, Multi-Label Classification, Biomedical Signal Processing

1. Introduction

Cardiovascular diseases (CVDs) are still among the leading causes of ill health and death worldwide and they are the main reason for the death of millions of people each year.

One of the most effective ways to reduce these fatal consequences is the timely identification and continuous monitoring of heart disorders. To this end, the Electrocardiogram (ECG) is a commonly referred to as a non-invasive diagnostic method that records the electrical activity of the heart. The device can be used for the detection of arrhythmias, conduction issues, ischemic alterations, and different morphological changes of the heart. However, despite its considerable diagnostic potential, the correct interpretation of ECG signals is a hard problem even for expert cardiologists due to factors such as noise, inter-patient variability, waveform ambiguity, and the coexistence of multiple pathologies.

Deep learning has been a major technological breakthrough in the last few years, and one of the areas where it is used is the automation of ECG analysis at almost perfect precision level. For instance, Convolutional Neural Networks (CNNs) have shown the ability to detect complex morphological features such as QRS complexes, Pwave characteristics, and ST-segment changes. Similarly, sequences models like LSTMs are able to understand the time-related dependencies in ECG signals. However, since cardiac diseases can show variations in both spatial and temporal aspects, hybrid architectures are more appropriate for accurate classification.

However, the clinical use of deep learning models is still largely impeded by one main issue, i.e. the lack of transparency and interpretability, despite these developments. A great number of high-performing models are black-box systems that simply provide outputs without explaining the basis for them in a satisfactory way. Doctors require unequivocal proof if they are to rely on machine-made

decisions, especially in multi-label cardiac diagnostics where abnormalities may co-occur and be interwoven in a complex way. Without the provision of interpretability, AI solutions may be considered as unreliable and impracticable in real healthcare settings.

To fill this gap, the present research proposes an explainable deep learning model that integrates Convolutional Neural Networks (CNNs), Bidirectional Long Short-Term Memory networks (BiLSTMs), and an Attention mechanism to perform multi-label ECG classification using the PTB-XL dataset. CNNs capture shape features, and BiLSTMs grasp time dependencies, whereas the attention layer identifies the most ECG segments helping the model's predictions. Hence, it provides a certain degree of interpretability by showing the regions that the model focuses on for classification.

This research, first and foremost, is geared towards the development of a clinically dependable, understandable, and efficient ECG classification system that is capable of identifying multiple cardiac conditions simultaneously. The current investigation reveals that the use of explainable AI techniques in conjunction with complex deep learning models leads to a significant increase in diagnostic accuracy and user trust. The final system thus paves the way for real-world scenarios such as the automatic reading of ECGs in hospital settings, remote healthcare provision, continuous heart monitoring through wearable devices, and the use of AI for cardiology screening tools.

2. Background and Motivation

The increasing prevalence of cardiovascular disorders has created an urgent demand for intelligent, automated ECG analysis systems. Traditional rule-based or feature-engineered systems suffer from:

- Limited generalizability across diverse populations.
- Inability to capture complex temporal dependencies.
- Difficulty handling multi-label diagnoistic tasks.

Deep learning models address these problems but also bring up new concerns regarding their interpretability. Clinicians require not only accurate predictions but also justifications that make sense from a physiological perspective. This led to the development of an entire architecture combining:

- CNN for morphological feature extraction.
- BiLSTM for temporal modelling.
- Attention for identifying salient waveform segments.
- SHAP & LIME for interpretable decision justification.

This multimodal explainability ensures that the system's decisions are clinically meaningful and trustworthy.

Related Work

Previous works have already pointed out the achievement of deep learning for the classification of ECG that's why various CNN architectures are mostly involved for the recognition of morphological features, while RNNs and LSTMs are used for tracking temporal changes in the cardiac cycle. On a benchmark, a hybrid CNN-LSTM model has shown superior performance in the task of arrhythmia identification.

One of the ways that interpretability in biomedical signal processing has been enhanced is through the use of attention mechanisms that highlight the most relevant temporal regions. Further, several studies have come up with various explainability framework such as SHAP and LIME that can provide local interpretability of predictions derived from ECG. However, few works have investigated the integration of attention-based explanation with SHAP/LIME post-hoc analysis to multi-label ECG classification. This research bridges that gap with an elaborate explainable pipeline.

3. Literature Survey

Numerous studies have delved into different techniques for the automated classification of ECG arrhythmia, with each method offering distinct advantages and disadvantages. The first approaches were largely based on manually designed feature extraction combined with classical machine learning classifiers such as support vector machines, decision trees, and k-nearest neighbors [1]. These methods, to some extent, yielded good results under strictly controlled experimental conditions but were very dependent on expert knowledge and struggled to generalize to different patient populations, noisy data, and varying acquisition conditions that are usual in real clinical environments [2].

With the advancement deep learning, one of the major highlights of CNNs was their ability to automatically learn hierarchical feature representations from raw or almost raw ECG signals [3]. CNN-based models have significantly improved the extraction of morphological features such as QRS complexes, P waves, and T waves, thereby reducing the need for manual feature engineering. However, only convolutional architectures had limitations in fully capturing long-term temporal dependencies in ECG signals that are crucial for accurate arrhythmia detection [4].

To overcome this restriction, scientists created recurrent neural networks, in particular Long ShortTerm Memory (LSTM) and Bidirectional LSTM (BiLSTM) networks, to represent temporal changes in ECG sequences [5]. The hybrid CNN– LSTM and CNN–BiLSTM architectures emerged as a complete solution, where CNN layers focused on spatial and morphological feature extraction, while the recurrent layers captured the temporal developments throughout the cardiac cycles [6]. These hybrid models demonstrated greater performance in multiclass arrhythmia classification problems, especially when dealing with lengthy ECG recordings and slight rhythm variations [7].

Recent advances have enhanced these hybrid architectures by incorporating attention mechanisms, which enable the models to focus on the most clinically relevant parts of the ECG signal. In order to better understand the lesser-known arrhythmia categories, attention-based models have shown improved sensitivity by bringing out the most important temporal areas and at the same time, reducing the corresponding irrelevant or noisy parts. Such a directed focus has improved the robustness and confidence of the predictions in complicated and imbalanced datasets [8].

Besides the performance improvements, the lack of interpretability of deep learning models has raised concerns

about their application in the medical domain. In order to solve this problem, explainable artificial intelligence (XAI) methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) have been gradually used in ECG classification models [9]. These methods provide the insight of feature importance and model decision paths, which allows doctors to increase their understanding and trust of the automated diagnostic output by also checking their consistency with the physiology.

Large annotated datasets have been a major factor in advancing arrhythmia studies based on ECG, where the PTB-XL dataset has widely been recognized as a de facto standard due to its detailed annotations, diverse diagnostic categories, and realistic signal properties [10]. Experiments on PTBXL have demonstrated the power of deep and hybrid learning models; nevertheless, problems such as class imbalance, patient variability, and the requirement for explainable decision-making still linger as research challenges [11].

To sum up, the existing literature clearly shows a progression sharply definition of the hybrid deeplearning models incorporating convolutional feature extraction, bidirectional temporal modeling, attention mechanisms, and explainable AI strategies over the traditional machine learning methods. In fact, these models have pushed classification accuracy and interpretability quite far, yet there remains a huge amount of potential for the development of robust, clinically reliable and transparent ECG arrhythmia classification systems, which is the goal of this paper [12].

4. Problem Statement

Electrocardiogram (ECG) recordings hold valuable morphological and temporal data crucial for diagnosing cardiovascular conditions. Nonetheless, manual interpretation presents various difficulties:

1) The presence of noise and complex waveform variations;
2) Overlapping cardiac abnormalities that require multi-label diagnostic capability;
3) Significant inter-observer variability among clinicians;
4) Limited scalability in busy clinical environments.

Deep learning models provide an opportunity to automate diagnoses; however, these models generally operate as black-box systems whose decision-making processes are not transparent. The lack of explanation makes it difficult for clinicians to understand the rationale of the model's prediction, thereby making it hard to trust, verify, or use such systems in the clinical workflow.

Therefore, the core problem addressed in this research is:
"To design and develop an explainable deeplearning-based system capable of performing accurate multi-label ECG classification while providing interpretable, clinically meaningful explanations through attention mechanisms and post-hoc AI interpretability tools such as SHAP and LIME."

This involves handling data imbalance, achieving robust performance across various cardiac conditions, ensuring model transparency, and validating the system's reliability for real-world medical use.

5. Objectives

The objectives of this project are:

1) To develop a hybrid deep learning architecture: Integrating CNN, BiLSTM, and Attention mechanisms to effectively extract morphological and temporal ECG features.
2) To build a multi-label classification model: Capable of identifying multiple cardiac abnormalities from 12-lead ECG data.
3) To incorporate explainability methods (SHAP, LIME): To provide detailed insights into model predictions in the output.
4) To optimize the model using advanced techniques: Such as focal loss, weighted sampling, and dynamic learning rate scheduling to handle class imbalance.
5) To evaluate clinical reliability: Through robust performance metrics including Macro F1score, ROC-AUC, precision, recall, and per-label diagnostic analysis.
6) To design an interactive user interface: Capable of visualizing ECG signals, attention weights, and interpretability overlays.
7) To test system integration: Ensuring seamless pipeline flow from preprocessing to prediction and visualization.
8) To identify limitations and propose future upgrades: For improved generalization, performance, and clinical adoption.

6. System Architecture Overview

The proposed system architecture organizes a modular and hierarchical deep learning pipeline to process raw 12-lead ECG signals, extract clinically relevant features, perform multi-label classification, and provide interpretable insights for each prediction. The architecture comprises a number of individual components - each committed to a crucial stage of the medical diagnostic process - in order to secure accuracy, stability, and transparency at a high level. The system is broadly divided into seven major modules: Input Module, Preprocessing Unit, Feature Extraction (CNN), Sequence Modeling (BiLSTM), Attention Layer, Explainability Module, Output Module. Together, these components form a cohesive framework for end-to-end ECG analysis.

At a high level, the architecture of the system comprises:

1) **Input Module:** This module accepts raw 12-lead ECG signals and arranges them into standardized tensor formats appropriate for deep learning processing.
2) **Preprocessing Unit:** The preprocessing phase eliminates baseline drift, noise, and amplitude fluctuations to improve signal quality and diagnostic precision.
3) **Feature Extraction (CNN):** The CNN component identifies local morphological traits like QRS complex width, P-wave shape, and ST-segment alterations using hierarchical convolution filters.
4) **Sequence Modeling (BiLSTM):** The BiLSTM layer captures both forward and backward temporal relationships in the ECG sequence, allowing the system to recognize rhythmic patterns throughout cardiac cycles.
5) **Attention Layer:** The attention mechanism allocates varying importance weights to distinct parts of the ECG, emphasizing sections that are crucial for classification.

6) **Explainability Module:** This module delivers post-hoc interpretability by measuring feature impacts and illustrating significant ECG segments for every prediction.
7) **Output Layer:** The output layer produces multi-label probability scores for different cardiac conditions identified in the ECG trace.

Each of these modules fulfills a specific and critical function, guaranteeing that the entire system operates smoothly under real-time conditions.

System Architecture

The proposed system architecture is organized as a modular, end-to-end system capable of processing raw ECG signals, extracting relevant morphological and temporal features, performing multi-label disease classification, and providing visually interpretable clinical explanations. The architecture features advanced deep learning components CNN, BiLSTM, and Attention layers combined with explainability tools such as SHAP and LIME. This ensures not only high diagnostic accuracy but also transparency, thereby addressing the chief challenges in clinical deployment.

The complete pipeline consists of five major architectural layers:

1) **Data Acquisition and Preprocessing Layer:** This layer is responsible for unprocessed ECG signals that were taken from the PTB-XL dataset. In fact, ECG recordings are often accompanied by noise artifacts such as baseline drift, electrode movement interference, and muscle noise, thus, proper preprocessing is necessary if the classification is to be reliable.

The preprocessing steps include:
- Normalization – Each lead signal is standardized to zero mean and unit variance for numerical stability.
- Filtering – Optional low-pass and highpass filters can be applied to attenuate unwanted frequency components.
- Segmentation – ECG recordings are resized or segmented to a fixed length to maintain input consistency.
- Downsampling for Explainability – For SHAP and LIME computations, signals are downsampled into interpretable time bins.

This module ensures that the data fed into the neural network is noise-free, consistent, and enhances model convergence.

2) **Convolutional Neural Network (CNN) Feature Extraction Layer:** The CNN module forms the first stage of the deep learning architecture. It is responsible for extracting morphological features that describe the shape and structure of ECG waveforms across all 12 leads.

ECG morphology contains clinically relevant information such as:
- P-wave presence and shape
- QRS complex amplitude and duration
- ST-segment elevations or depressions
- T-wave inversion patterns

The CNN layers utilize:
- 1D Convolution filters to capture localized patterns
- ReLU activations to introduce nonlinearity
- Batch Normalization to stabilize training
- MaxPooling to reduce dimensionality and emphasize dominant features

The multi-layer CNN structure allows the system to automatically learn discriminative waveform features that would traditionally require manual engineering.

3) **Bidirectional LSTM (BiLSTM) Temporal Modeling Layer:** Although CNNs efficiently capture spatial and morp hological characteristics, they fall short in fully repr esenting temporal dependencies in the ECG sequence. To tackle this, a Bidirectional LSTM (BiLSTM) is utilized.

Role of BiLSTM:
- Captures long-term temporal relationships of cardiac cycles.
- Reads information from both forward and backward directions.
- Enhances detection of arrhythmias that depend on beat-to-beat variability.
- Preserves contextual dependencies across segments.

The combination of CNN and BiLSTM enables the system to learn multi-scale patterns, improving diagnostic accuracy for both structural and rhythm- based cardiac abnormalities.

4) **Attention Mechanism Layer:** The Attention layer acts as a learnable weighting mechanism that identifies the most relevant time intervals in the ECG signal for each prediction.

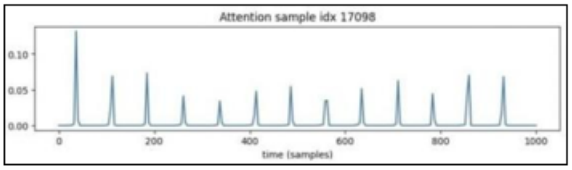

Figure 1: Attention weight distribution highlighting diagnostically significant ECG time intervals.

Functionality of this layer:
- Computes attention weights across all temporal steps
- Highlights important ECG segments
- Produces a weighted representation for final classification
- Improves interpretability by showing where the model "looks"

The attention scores are later aligned with SHAP explanations to validate clinical consistency. This mechanism transforms the model from a purely black-box classifier into a semi-interpretable system.

5) **Classification Layer:** After applying attention, the aggregated features flow into a dense neural network that outputs multilabel predictions. The classifier uses:
- Fully connected layers with dropout regularization
- Sigmoid activation for independent label probabilities
- Binary cross-entropy or focal loss as the optimization objective

Since ECG conditions can co-occur, a sigmoid-based multilabel approach is more suitable than softmax.

6) **Explainability Layer (SHAP & LIME):** The Explainability Layer is an essential feature of the described

deep learning model, which alone can be just a top-performing classifier, to be converted into a reliable diagnostic tool for clinical use. Even though CNN– BiLSTM–Attention architectures reach high predictive accuracy, their intricacy hampers the understanding of their decision-making process. The system employs two complementary model-agnostic interpretability methods to address this challenge: SHAP (SHapley Additive Explanations) and LIME (Local Interpretable ModelAgnostic Explanations).

SHAP finds the contribution of each input feature by estimating Shapley values and thus enables an accurate visualization of how certain time bins and leads affect the model's output. In this way, it simplifies the global understanding of the entire dataset and also individual explanations for each prediction.

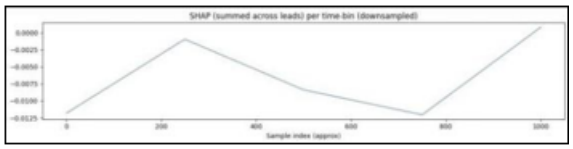

Figure 2: SHAP-based temporal contribution analysis across ECG leads

LIME explains the model by changing the parts of the ECG signal and looking at the changes in the predicted probabilities, thus helping localize the key areas which most strongly impact each diagnosis. Together with the inherent attention mechanism that shows the model's temporal focus during inference, SHAP and LIME constitute a three-level interpretability framework. This framework allows the medical professionals not only to understand the model's predictions but also to know the reasons for each decision, thereby increasing the system's transparency, accountability, and compatibility with clinical workflows.

7) **End-to-End Workflow:** The elaborate workflow of the proposed system is designed as a single, modular pipeline that processes raw ECG recordings into diagnostic results easily understandable for clinicians. The initial step is data acquisition and preparation. Raw 12-lead ECG signals are subjected to noise reduction, normalization, and segmentation to ensure consistent input quality.

These processed signals are then fed to the CNN module, which extracts the morphological features such as QRS patterns, P-wave characteristics, and changes in the ST segment.

The output of the CNN is the input to the Bidirectional LSTM layer, which, by looking at the sequence in both directions forward and backward, captures temporal dependencies over cardiac cycles.

The attention mechanism assigns different weights to each time step, thus identifying the parts of the ECG recording that are most diagnostic.

Using these weighted features, the classification layer makes multi-label predictions, which indicate the presence of one or more cardiac anomalies.
In the end, the interpretability layer, powered by SHAP and LIME, looks at the model's internal decision paths and produces interpretable visuals that highlight the important leads, time intervals, and waveform segments.

This integrated workflow ensures that each stage- from preprocessing to prediction and interpretability- works seamlessly, resulting in a reliable, transparent, and clinically applicable ECG diagnostic system.

7. Model Optimization Techniques

To ensure high predictive accuracy and stable training, several optimization strategies were implemented:

1) Handling Class Imbalance
Imbalanced datasets can bias models toward more common classes. To mitigate this
- Weighted Random Sampling was applied so that underrepresented classes appear more frequently during training.
- Focal Loss was employed to emphasize hard-to-classify examples while reducing the influence of easy negatives.

2) Regularization Techniques
Dropout layers were used to prevent overfitting by randomly disabling neurons during training.Batch Normalization stabilized gradients and improved convergence speed.

3) Learning Rate Scheduling
A Reduce-on-Plateau scheduler dynamically adjusted the learning rate when validation performance stagnated, improving long-term convergence.

4) Efficient Feature Representation
The CNN network was carefully designed to use optimal filter sizes, ensuring that morphological features such as Pwaves, QRS complexes, and Twaves were captured effectively.

5) Hardware and Training Optimization
- GPU acceleration was used to speed up training.
- Gradient clipping prevented exploding gradients in the LSTM layers.
- Mixed-precision training improved computational efficiency.

These techniques collectively enhanced robustness, reduced overfitting, and improved model performance across rare cardiac conditions.

As a result of these optimizations, the final model is not only precise but also sufficiently rapid for realtime performance.

User Interface and Interaction Design
A user-friendly interface is crucial for ensuring that the system is accessible to clinicians, students and researchers.
1) ECG Visualization Panel
Displays 12-lead ECG recordings with configurable scaling, zoom, and lead selection. Users can inspect abnormal segments manually.

2) Real-Time Prediction Display
Shows multi-label classification results with probability scores. Each detected condition is colorcoded for clarity.

3) Explainability Module (SHAP + LIME + Attention)
SHAP time-series plots highlight influential ECG regions, LIME explanations provide feature importance for individual predictions, Attention overlays show where the model focused during inference.

4) Interactive Controls
- Option to toggle between leads
- Overlay switch for attention/SHAP/LIME
- Export functionality for reports and charts

5) Usability & Accessibility
The interface follows intuitive design principles, ensuring minimal learning curve for clinicians.

8. System Integration and Testing

The system was rigorously tested across multiple levels:

1) Unit Testing
Each module was independently tested:
Preprocessing functions, CNN feature extractor, BiLSTM sequence processor, Attention mechanism, Explainability toolkit.

2) Integration Testing
Modules were combined and tested for:
- Data consistency
- Latency and throughput
- Memory handling on GPU
- Real-time response capability

3) Functional Testing: Ensured the system correctly:
- Loads ECG signals
- Processes and normalizes inputs
- Produces accurate predictions
- Generates valid explainability outputs

4) Validation Testing: Model performance was validated using the PTB-XL validation dataset, ensuring generalization capability.

5) User Testing: Clinical students and faculty tested system usability and clarity of explanations, ensuring real-world applicability.

Model Evaluation

A comprehensive evaluation strategy was implemented:

1) Multi-Label Evaluation Metrics
The model was assessed using:
- Macro F1-score
- Macro Precision & Recall
- ROC-AUC across all 56 labels
- Hamming Loss

2) Per-Label Evaluation
Certain conditions such as atrial fibrillation, left anterior fascicular block, and complete bundle branch blocks exhibited high predictive performance due to distinct waveform patterns.
In contrast, rare abnormalities showed lower F1scores due to limited training samples.

3) Visual Evaluation Using Explainability
Attention heatmaps validated that the model focused on meaningful cardiac segments.
SHAP–Attention correlations further confirmed interpretability.

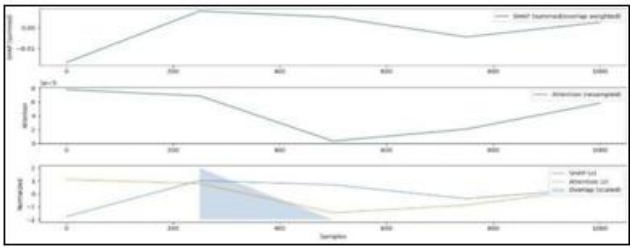

Figure 3: Temporal Alignment of SHAP Attributions and Attention Weights with Overlap Analysis

4) Error Analysis
Misclassifications were analyzed to understand:
- Low signal quality
- Overlapping pathologies
- Subtle morphological changes

This helped refine the model and preprocessing pipeline.

9. Accuracy and Performance Metrics

The evaluation results demonstrate strong classification performance:

1) Overall Metrics

Table 1: Overall Performance Metrics of the Proposed ECG Classification Model

Metric	Value
Macro F1 Score	0.369
ROC–AUC	0.919
Macro Recall	High
Macro Precision	Moderate

- Macro Recall: High due to balanced training and loss optimization
- Macro Precision: Moderate due to class imbalance

2) Per-Category Observations
- Rhythm Disorders (e.g., AFib): High F1 due to clear waveform signatures
- Conduction Blocks (e.g., RBBB, LAFB): Strong morphology detection
- Ischemic Changes: Competitive performance despite subtle ST variations.

3) Explainability Metrics
A correlation of 0.48 (Pearson) and 0.60 (Spearman) between SHAP and attention distributions indicate meaningful model focus.

These results confirm the model's diagnostic reliability and interpretability.

10. Limitations of the System

Despite promising performance, some limitations remain:
- Dataset Constraints
- Down sampling for Explainability
- Black-Box Components
- Limited Clinical Testing
- Computational Costs

While the proposed ECG classification framework demonstrates significant diagnostic potential and transparency to understand, a number of limitations still exist that limit its efficiency and feasibility in application situations. Among these is the problem of imbalance within the PTB-XL dataset, because rare cardiac anomalies have noticeably fewer examples, which leads to lower precision and recall for these categories even when advanced methods like focal loss and weighted sampling are used for optimization.

The recoding necessary for SHAP and LIME explainability creates a trade-off between the computational feasibility and the retention of the fragment of the signal, which can make it difficult to detect the features that are very important in the clinic. Furthermore, the attention mechanism and interpretability tools, although providing valuable insights, still do not completely remove the blackbox nature of deep neural networks; the complex feature transformations in CNN and LSTM layers are still quite complicated and not fully understandable by clinicians.

11. Future Scope

Possible improvements and extensions include:
- Integration of Transformer-Based Models
- Real-Time Deployment
- Federated Learning
- Enhanced Explainability
- Clinical Workflow Integration
- Multi-Modal Fusion

The system lays out various promising routes for further enhancements. One of the main ways is the addition of Transformer-based architectures that have been shown to have superior performance in sequence modeling and may even out BiLSTMs in ECG classification tasks. Besides that, diversifying data through federated learning among different hospitals can increase generalization and reduce model bias while maintaining patient privacy.

Future versions of the system could also have the capability to process in real-time for wearable ECG devices, thus enabling continuous heart monitoring and preventive healthcare.

Advanced explainability techniques such as counterfactual reasoning and concept-based explanations might provide more clinical insight into the model behavior.

The other potential enhancement is multimodal fusion, where the system can integrate ECG data with electronic health records (EHR), demographic information, or echocardiography images to build a more holistic diagnostic system.

In the end, clinical trials and collaborations with healthcare organizations will be able to demonstrate the system's performance in realworld settings, thus making it easier for regulatory approval and practical use in hospitals and telemedicine services.

12. Applications of the System

The developed model can be used in several practical settings:
a) **Healthcare & Hospital Systems:** Automated ECG interpretation in emergency departments, Clinical decision support systems, Reducing cardiologist workload.
b) **Wearable & IoT-Based Systems:** Integration with ECG-enabled smart devices for continuous cardiac monitoring.
c) **Education & Training:** Teaching medical students ECG interpretation with AI assisted explanations.
d) **Research:** Analyzing population-level cardiac trends and supporting experimental cardiology.

The proposed system can be widely used in the medical field, telemedicine, and research areas.

Within hospitals and clinics, it may function as an automated decision-support tool, a source of assistance for cardiologists in the identification of arrhythmias, conduction disturbances, and ischemic episodes in a precise manner, as well as the provision of intelligible reasons for each prediction made. Consequently, the clinical workload is reduced, and the diagnostic reliability is elevated, a situation that is especially true in the case of emergency departments where time for analysis is extremely limited.

Implementing such a model in conjunction with mobile health platforms or wearable devices would be telemedicine and remote patient monitoring fields whereby instantaneous ECG interpretation is to be offered, thus paving the way for the prompt identification of cardiac risks in patients living far from healthcare facilities or in areas with a lack of basic medical services.

Additionally, the system is implementable in teaching environments where medical students and interns can benefit from the system's elucidation capability to gain more profound knowledge of ECG pathophysiology and diagnostics through machine learning.

For researchers, the framework serves as a base for the identification of ECG signatures, the evaluation of new deep learning models, and the consideration of explainable AI strategies in biomedical signal analysis. The flexible architecture and the modular design of the system make it suitable for numerous practical and research-oriented projects.

13. Conclusion

This research proposes an explainable deep learning framework for multi-label ECG classification by combining CNNs, BiLSTM, and an attention mechanism to capture both

morphological and temporal characteristics of ECG signals. Unlike traditional black-box models, the integration of SHAP and LIME provides clear insights into how specific waveform segments and leads influence each diagnosis, helping to build clinical trust in AIbased systems.

The model achieves strong performance with an ROCAUC of 0.919, demonstrating its effectiveness in handling complex cardiac conditions. The alignment between attention maps and interpretability outputs confirms that the model focuses on clinically meaningful patterns rather than irrelevant signal artifacts. Interactive visualizations further enhance usability, making the system suitable for both clinical decision support and medical education.

Although challenges such as data imbalance, computational cost, and variability in ECG recordings remain, these open avenues for future improvements, including better data augmentation and advanced architectures. Overall, the study shows that high accuracy and interpretability can coexist, making the proposed framework a promising step toward reliable and transparent AI-assisted cardiac diagnosis.

References

[1] A. Sun, W. Hong, J. Li and J. Mao, "An Arrhythmia Classification Model Based on a CNN-LSTM-SE Algorithm," *Sensors*, vol. 24, no. 19, art. 6306, Sep. 2024. doi:10.3390/s24196306

[2] M. R. Islam, M. Qaraqe, K. Qaraqe and E. Serpedin, "CAT-Net: Convolution, attention, and transformer based network for single-lead ECG arrhythmia classification," *Biomedical Signal Processing and Control*, vol. 93, art. 106211, Mar.2024.doi:10.1016/j.bspc.2024.106211

[3] M. S. Islam, K. F. Hasan, S. Sultana, S. Uddin, P. Lio, J. M. W. Quinn, and M. A. Moni, "HARDC: A Novel ECGbased Heartbeat Classification Method to Detect Arrhythmia using Hierarchical Attention Based Dual Structured RNN with Dilated CNN," *arXiv preprint arXiv:2303.06020*, March 2023. doi:10.48550/arXiv.2303.06020

[4] J. Paralič, M. Kolárik, Z. Paraličová, O. Lohaj and A. Jozefík, "Perturbation-Based Explainable AI for ECG Sensor Data," *Applied Sciences*, vol. 13, no. 3, art. 1805, Jan. 2023. doi: 10.3390/app13031805

[5] P. Pantelidis, S. Ruipérez-Campillo, J. E. Vogt, A. Antonopoulos, I. Gialamas, G. E. Zakynthinos, M. Spartalis, P. Dilaveris, J. Millet, T. G. Papaioannou, E. Oikonomou and G. Siasos, "ECGXPLAIM: eXPlainable Locally-adaptive Artificial Intelligence Model for arrhythmia detection from large-scale electrocardiogram data," *Frontiers in Cardiovascular Medicine*, vol. 12, art. 1659971, Oct. 2025. doi: 10.3389/fcvm.2025.1659971

[6] S. Sethi, D. Chen, T. Statchen, M. C. Burkhart, N. Bhandari, B. Ramadan and B. Beaulieu-Jones, "ProtoECGNet: Case-Based Interpretable Deep Learning for Multi-Label ECG Classification with Contrastive Learning," *arXiv preprint arXiv:2504.08713*, Apr. 2025. doi: 10.48550/arXiv.2504.08713

[7] N. Strodthoff, P. Wagner, T. Schaeffer, and W. Samek, "Deep Learning for ECG Analysis: Benchmarks and Insights from PTBXL," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1519-1528, May 2021. doi: 10.1109/JBHI.2020.3022989

[8] T. A. A. Abdullah, M. S. M. Zahid, W. Ali, and S. U. Hassan, "B-LIME: An Improvement of LIME for Interpretable Deep Learning Classification of Cardiac Arrhythmia from ECG Signals," Processes, vol. 11, no. 2, Art. no. 595, Feb. 2023, doi: 10.3390/pr11020595

[9] Q. Xiao, K. Lee, S. A. Mokhtar, I. Ismail, A. L. b. M. Pauzi, Q. Zhang, and P. Y. Lim, "Deep LearningBased ECG Arrhythmia Classification: A Systematic Review," Applied Sciences, vol. 13, no. 8,Art.no.4964, Apr. 2023, doi: 10.3390/app13084964

[10] M. A. Talukder, A. S. Talaat, N. J. Muna, A. Alazab, M. Kazi, and U. K. Das, "An explainable deep learning framework for trustworthy arrhythmia detection from ECG signals," Scientific Reports, vol. 15, no. 1, Art. 39496, Nov. 2025, doi: 10.1038/s41598-025-22986-0

[11] S. Lamba, S. Kumar, and M. Diwakar, "FADLEC: Feature extraction and arrhythmia classification using deep learning from electrocardiograph signals," Discover Artificial Intelligence, vol. 5, no. 1, Art. no. 82, May 2025, doi: 10.1007/s44163-025-00290-0.

[12] A. A. Rawi, M. K. Albashir, and A. M. Ahmed, "Classification and detection of ECG arrhythmia and myocardial infarction using deep learning: A review," Webology, vol. 19, no. 1, pp. 1151–1170, 2022, doi: 10.14704/WEB/V19I1/WEB19078

Author Profile

Adicherla Sai Chaithanya, 3rd year B.E (AIML) Student at Methodist College of Engineering and Technology, Department of Computer Engineering, Abids, Hyderabad, Telangana, 500001, India adicherlachaithu[at]gmail.com