Explainable AI-Powered ECG Anomaly Detection Using SHAP and LSTM Autoencoders

K. Nayana
Department of Computer Science and Engineering, GMIT Davanagere, Visvesvaraya Technological University, Belagavi, Karnataka, India
nayana.k.research@gmail.com (corresponding author)

S. Vinay
Department of Computer Science and Engineering, PESCE Mandya, Visvesvaraya Technological University, Belagavi, Karnataka, India
vinaymanyan@gmail.com

Received: 23 January 2026 | Revised: 7 March 2026 | Accepted: 15 March 2026

 DOI: https://doi.org/10.48084/etasr.17734

ABSTRACT

Cardiovascular diseases are the leading cause of mortality, underscoring the need for accurate and early detection of cardiac abnormalities. This study proposes a Long Short-Term Memory (LSTM) autoencoder framework for automated Electrocardiogram (ECG) anomaly detection, integrated with SHapley Additive exPlanations (SHAP) to ensure transparent and clinically interpretable decision-making. Using the ECG5000 dataset, the model learns the intrinsic characteristics of normal cardiac cycles and identifies deviations through reconstruction error analysis. The proposed framework achieved an accuracy of 97.6%, a precision of 95.8%, and an F1-score 96.7%, outperforming traditional machine-learning baselines and earlier deep-learning approaches. SHAP-based visualization highlights the specific temporal segments influencing anomaly predictions, enhancing clinical trust and applicability. This work demonstrates a robust, explainable, and efficient approach, suitable for real-time cardiac monitoring.

Keywords-ECG anomaly detection; LSTM autoencoders; explainable AI; SHAP

I. INTRODUCTION

Cardiovascular diseases are the main cause of mortality, with a significant increase in death rates driven by demographic aging and lifestyle-related risk factors being reported. Transformer-based and hybrid deep learning architectures have been explored to address this issue by enabling effective Electrocardiogram (ECG) anomaly detection under noisy and heterogeneous clinical conditions [1]. In addition to these advances, automated ECG analysis has gained prominence due to its potential to reduce diagnostic delays and improve early detection of life-threatening cardiac events.

ECG is the most widely adopted non-invasive technique for cardiac assessment, offering real-time insights into cardiac electrical activity. However, conventional ECG interpretation methods face challenges in scalability and consistency, particularly when dealing with long-term monitoring data. Recent deep learning frameworks have demonstrated that end-to-end models can learn discriminative temporal features directly from raw ECG signals, significantly improving detection accuracy compared to traditional approaches [2]. In particular, unsupervised learning approaches have gained attention for reducing dependency on large annotated datasets.

Among these approaches, Long Short-Term Memory (LSTM)-based autoencoders have proven effective in modeling long-term temporal dependencies and learning intrinsic representations of normal cardiac rhythms. By minimizing reconstruction error on healthy ECG segments, such models can identify anomalous deviations without explicit anomaly labels, as demonstrated in recent ECG autoencoder frameworks [3]. Despite their effectiveness, these deep learning models often lack transparency, making it difficult for clinicians to understand the basis of anomaly predictions. To address this limitation, Explainable Artificial Intelligence (XAI) is a critical requirement in medical AI systems. Techniques such as SHapley Additive exPlanations (SHAP) enable the attribution of model outputs to individual input features or time steps, thereby enhancing clinical interpretability and trust in automated ECG analysis systems [4].

Despite advances in automated ECG analysis, clinical practice continues to rely on manual interpretation by trained cardiologists, limiting scalability and increasing diagnostic

workload. Feature-based machine learning techniques, such as K-Nearest Neighbors (KNN), have been employed for ECG classification; however, their performance is constrained by reliance on engineered features, sensitivity to noise, and inter-patient variability [5]. These limitations hinder their applicability in continuous monitoring environments.

Ensemble-based approaches, including Random Forest classifiers, have improved robustness by leveraging multiple decision trees, yet they still depend on engineered features and struggle to generalize across diverse ECG morphologies. While such models improve heartbeat classification accuracy, they fail to capture complex temporal dependencies inherent in ECG signals [6]. Consequently, their effectiveness diminishes when applied to real-world clinical datasets. Support Vector Machine (SVM)–based ECG classification methods have also been explored, particularly for shockable rhythm detection. Although SVMs demonstrate reasonable accuracy, their performance degrades in highly imbalanced datasets and dynamic physiological conditions, limiting their reliability for anomaly detection tasks [7]. These shortcomings underscore the need for more adaptive and temporally aware models. Comparative analyses of traditional machine learning and deep learning approaches reveal that while deep models outperform classical techniques, challenges related to computational complexity, interpretability, and real-time deployment persist. Such limitations restrict their integration into wearable and remote healthcare systems, where efficiency and transparency are crucial [8].

Research in ECG anomaly detection has focused on sophisticated deep learning architectures capable of modeling complex temporal dependencies and handling real-world signal variability. Transformer-based hybrid models, such as Deep ECG-Net, have demonstrated efficiency by leveraging attention mechanisms to improve anomaly detection performance under noisy and heterogeneous ECG conditions [1]. Similarly, deep learning frameworks for automated obstructive ECG analysis have shown that end-to-end learning significantly enhances diagnostic accuracy compared to conventional pipelines based on handcrafted features [2].

Unsupervised deep learning approaches have gained prominence due to their reduced dependence on labeled data. In this regard, authors in [3] proposed ECG-NET, a three-layer LSTM autoencoder that learns intrinsic patterns of normal ECG signals and identifies anomalies through reconstruction error, effectively mitigating overfitting and noise sensitivity in long-term recordings. Despite their effectiveness, such deep models often operate as black boxes, limiting their acceptance in clinical practice. To address interpretability concerns, explainable deep learning is an significant research area. Authors in [4] highlighted the importance of integrating XAI techniques, such as SHAP and LIME, into ECG analysis frameworks, enabling transparent attribution of model predictions to specific signal segments and improving clinician trust.

Alongside deep learning, classical machine learning methods are also explored for ECG analysis. Optimized KNN-based approaches have been proposed for ECG classification; however, their reliance on distance-based metrics and handcrafted features restricts scalability across diverse patient populations [5]. Random Forest–based heartbeat classification methods improve robustness through ensemble learning but are limited in capturing long-term temporal dependencies inherent in ECG signals [6]. Similarly, SVM-based ECG classification methods incorporating advanced feature extraction have shown effectiveness in detecting shockable rhythms, though their performance remains sensitive to feature quality and class imbalance [7]. Comprehensive comparative studies evaluating traditional machine learning and deep learning models consistently demonstrate the superiority of deep learning approaches for ECG anomaly detection, particularly in complex and imbalanced datasets, although challenges related to computational cost and interpretability persist [8]. Class imbalance, in particular, is a significant issue in ECG datasets, as abnormal cardiac events occur far less frequently than normal rhythms, leading to biased predictions if not properly addressed [9]. To enhance discriminative capability, hybrid approaches combining Convolutional Neural Networks (CNNs) with SVM classifiers using ECG spectrogram representations have been proposed, effectively capturing both spatial and frequency-domain features for improved arrhythmia classification [10]. Tree-based ensemble learning techniques, such as Extreme Gradient Boosting, have also been applied to ECG signal classification, offering improved accuracy through boosted decision trees, albeit with limited temporal modeling capability due to feature dependency [11].

Furthermore, hybrid deep learning architectures integrating CNN and LSTM components have been introduced, enabling joint modeling of morphological and temporal ECG characteristics and achieving superior performance in cardiovascular disease diagnosis [12]. As model complexity increases, the need for explainability becomes crucial, with SHAP-based XAI frameworks demonstrating their effectiveness in enhancing transparency and clinical trust in automated cardiac diagnosis systems [13]. The availability of standardized benchmark datasets has further accelerated progress, with the ECG5000 dataset [14] serving as a widely used public resource for evaluating ECG anomaly detection models under reproducible experimental conditions.

II. METHODOLOGY

A. Dataset

This study utilizes the ECG5000 dataset [14], derived from a 20-h ECG recording obtained from the PhysioNet database of a patient with severe congestive heart failure. The dataset contains 5,000 extracted heartbeat signals, where each heartbeat is standardized to equal length using interpolation to ensure consistent input representation for computational models. Automated annotation methods assign class labels corresponding to different heartbeat patterns present in the recording. This structured dataset is deployed for benchmarking time-series classification and anomaly detection models. It enables reliable training and evaluation of machine learning and deep learning approaches for ECG analysis. By learning patterns of normal cardiac rhythms, the model can effectively detect deviations associated with abnormal cardiac activity.

1) Participant Categorization

The dataset participants are divided into two defined cohorts: Normal ECG Group and Anomalous ECG Group. The Normal ECG Group cohort consists of individuals exhibiting normal sinus rhythm without any clinically documented cardiac irregularities. The recordings from this group serve as the baseline in the training paradigm, allowing the LSTM autoencoder to learn the intrinsic temporal and morphological properties of healthy cardiac cycles. The Anomalous ECG Group subjects are diagnosed with cardiac pathologies such as Atrial Fibrillation (AFib), Ventricular Tachycardia (VT), Premature Ventricular Contractions (PVC), and other arrhythmic or ischemic abnormalities. This diverse anomaly spectrum enables evaluation against real-world clinical variability, enhancing model robustness and reliability in detecting cardiac events.

2) Data Acquisition and Annotation Quality

The dataset acquisition process adhered to stringent clinical standards to ensure both signal quality and annotation precision. Raw ECG signals were obtained using industry-standard multi-lead devices offering high temporal resolution. Each recording was subjected to preprocessing, including noise filtering, baseline correction, and signal normalization to address common artifacts and inter-patient variability. Cardiologist annotations were cross-validated to minimize labeling errors, following recognized guidelines from the American Heart Association (AHA) and the Association for the Advancement of Medical Instrumentation (AAMI).

3) Preprocessing and Segmentation

For modeling, the continuous ECG waveforms were divided into fixed-length segments (windows) of 140 time-steps each. This window length balances the need for sufficient contextual signal information and computational efficiency during model training and inference. Amplitude normalization techniques standardize signal scales across subjects, compensating for electrode placement differences and individual physiological variations.

4) Data Splitting Strategy

The division into training, validation, and testing subsets was performed using stratified random sampling to preserve the distribution of normal and anomalous samples across splits. Approximately 80% of normal ECG segments comprise the training set, enabling the autoencoder to learn robust normative cardiac rhythms exclusively. The remaining 20%, containing both normal and anomalous samples, forms the test set to evaluate anomaly detection performance under realistic mixed conditions. Figure 1 shows the distribution of ECG samples across target classes, highlighting the dataset's class imbalance and motivating the use of reconstruction-based anomaly detection.

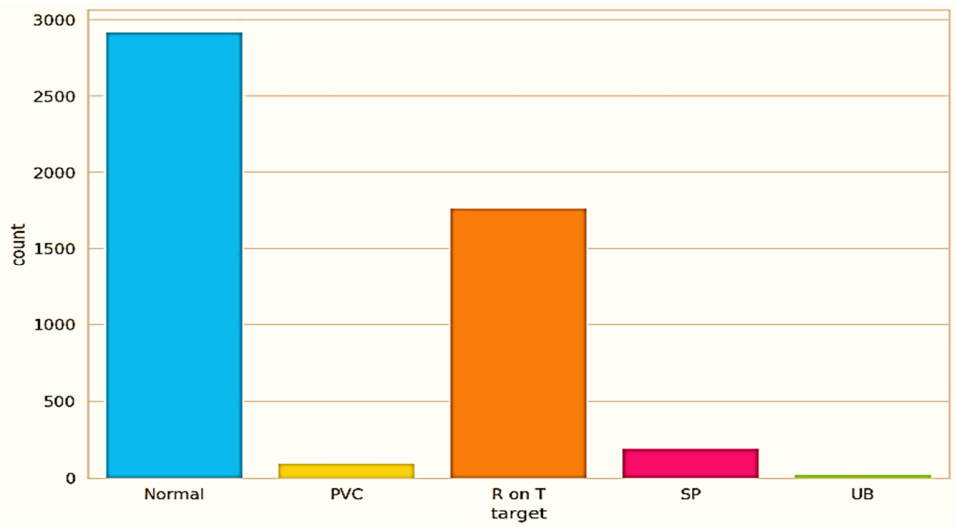

Fig. 1. Distribution of ECG samples across target classes.

5) Anonymization and Ethical Compliance

All identifying metadata were meticulously removed or anonymized to preserve patient privacy, aligning with ethical standards such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR). This process safeguards patient confidentiality and supports the wider sharing and reproducibility of research findings.

6) Integration with Explainability Framework

The rich clinical annotations and standardized segmentations enable seamless integration with the SHAP explainability pipeline. This synergy allows the model not only to detect anomalies but also to provide interpretable attributions at the individual time-step level, empowering clinicians with transparent insights into the decision-making process. In summary, the ECG dataset's patient diversity, rigorous annotation protocols, comprehensive preprocessing, and the strict data partitioning scheme provide an effective and

ethically sound foundation for developing and validating the proposed LSTM autoencoder and SHAP explainability framework. This approach is designed to maximize both predictive performance and clinical applicability in ECG anomaly detection.

B. ECG Signal Processing and Feature Engineering

An accurate and robust ECG anomaly detection system necessitates a comprehensive and meticulous signal preprocessing and feature engineering pipeline. Leveraging the code implementation and experimental visualizations, various methods were employed to enhance signal quality, preserve clinically relevant characteristics, and prepare data for the model. Figure 2 illustrates the mean ECG waveform for each heartbeat class, emphasizing temporal morphology differences between normal and abnormal patterns.

1) Data Normalization and Preprocessing

a) Min–Max Scaling

Raw ECG signals vary in amplitude due to physiological differences, sensor placement, and acquisition conditions. Each ECG segment was normalized to the [0,1] range using Min–Max scaling to ensure uniformity, equal feature contribution during training, and numerical stability in LSTM autoencoders, supporting faster convergence.

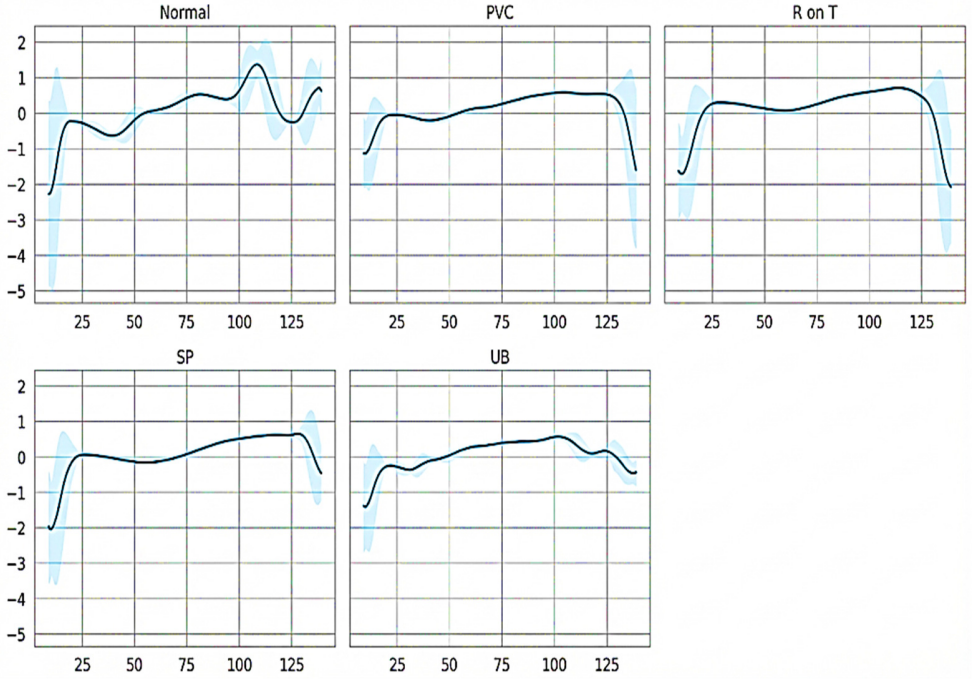

Fig. 2. Average ECG waveform of each class.

b) Reshaping to Fixed-Size Windows

Normalized ECG signals were segmented into fixed windows of 140-time steps and reshaped to (140,1) to match LSTM input requirements. This window size captures a complete cardiac cycle while minimizing computational overhead and avoiding fragmented representations.

c) Handling Class Imbalance with SMOTE

ECG datasets are typically imbalanced, with normal beats dominating abnormal cases. To enhance sensitivity to rare arrhythmic events, SMOTE is applied during validation and fine-tuning to generate synthetic minority samples and improve anomaly representation.

2) Segmentation and Feature Engineering

a) Segmentation into Pre-Defined Windows

ECG signals were divided into fixed-length windows of 140 samples to preserve temporal continuity and capture relevant local and global waveform characteristics.

b) Automated Feature Engineering via Deep Learning

Unlike engineered feature-based methods, the LSTM autoencoder learns spatiotemporal representations directly from raw ECG signals. This end-to-end approach captures short-term variations and long-term cardiac dynamics, enabling the detection of subtle anomalies beyond predefined clinical markers.

C. Deep Learning Model Architecture

The proposed autoencoder architecture was optimized for temporal ECG modeling and computational efficiency, as outlined below.

- Input Layer: The input layer was configured to accept normalized ECG windows of size (140,1), preserving sequential information.
- Encoder: The encoder was implemented using two stacked LSTM layers with 64 hidden units to compress the input

into a fixed-length latent representation and capture long-range dependencies and key heartbeat patterns.

- Decoder: The decoder was designed to mirror the encoder. LSTM layers built the ECG signal from the latent vector, restoring waveform characteristics of normal heartbeats.
- Output Layer and Loss: Reconstructed sequences pass through sigmoid activations, with Mean Absolute Error (MAE) used as the anomaly score, with higher errors indicating abnormalities.
- Hyperparameters and Training Procedures: Training was conducted using mini-batches of size 512, dropout regularization (0.2), the Adam optimizer, and early stopping after 20 epochs without validation improvement to enhance generalization and efficiency.

Figure 3 shows the reconstruction losses for training ECG data, illustrating normal signal reconstruction spread.

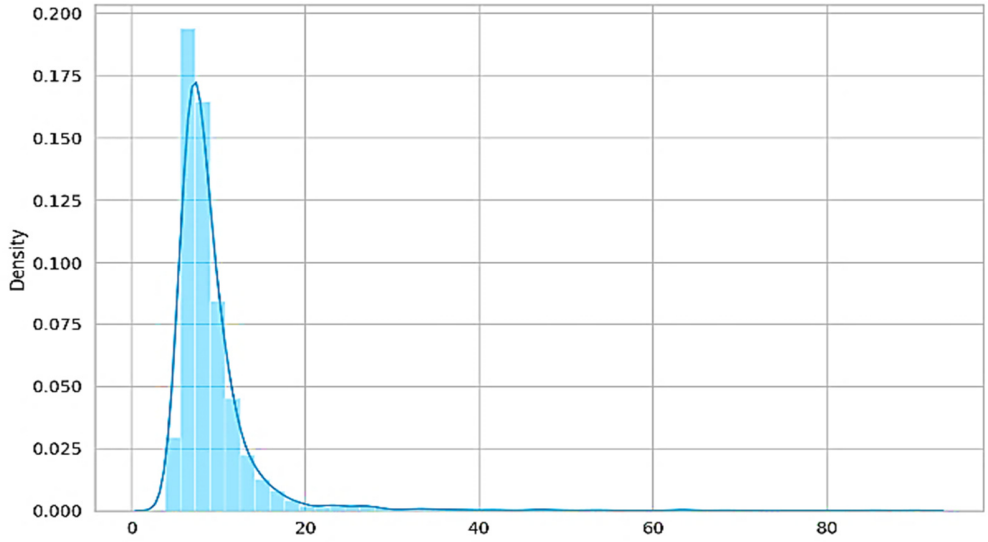

Fig. 3. Histogram of reconstruction losses for training ECG data.

D. Statistical Analysis and Performance Evaluation

Model performance was evaluated using accuracy, precision, recall, and F1-score to assess arrhythmia detection while balancing sensitivity and specificity. Training was monitored, and convergence was observed within 20–50 epochs, depending on data complexity. Normal signals exhibited low reconstruction errors, while anomalies showed pronounced spikes, enabling threshold-based classification. Overlay plots of original and reconstructed ECG signals were generated to highlight pathological morphological deviations. SHAP identified the time-step contributions driving reconstruction errors, with heat maps and force plots providing clinically interpretable insights, as shown in Figure 4. The integrated preprocessing, modeling, and explainability approach provided a framework for real-time ECG anomaly detection, applicable for clinical monitoring environments.

III. RESULTS AND DISCUSSION

A. Benchmarking Against Traditional Machine Learning

The performance of the proposed LSTM autoencoder was benchmarked against SVM and Logistic Regression (LR) models reported in [13] for ECG-based cardiac disease detection. The proposed model achieved an accuracy of 97.6%, a precision of 95.8%, and an F1-score of 96.7%, indicating reliable anomaly detection. The performance comparison is presented in Table I.

TABLE I. PERFORMANCE COMPARISON OF LSTM WITH TRADITIONAL MACHINE LEARNING MODELS

Metric	Proposed framework	LR model	SVM model
Accuracy	97.6%	85%	86%
Precision	95.8%	91%	89%
F1-score	96.7%	86%	88%

The superior performance of the proposed LSTM-based framework stems from its ability to learn long-term temporal dependencies directly from raw ECG signals, unlike SVM and LR models, which rely on handcrafted features. The reconstruction-based anomaly detection further improves separation between normal and abnormal patterns. Additionally, SHAP-based explainability enables transparent, time-step-level interpretation of predictions, a capability absent in traditional models. Overall, the proposed framework offers a more accurate, robust, and clinically interpretable solution. Figure 4 shows the contribution of individual time steps to the reconstruction error for a single ECG sample, while Figure 5 displays the distribution of time-step contributions across all samples.

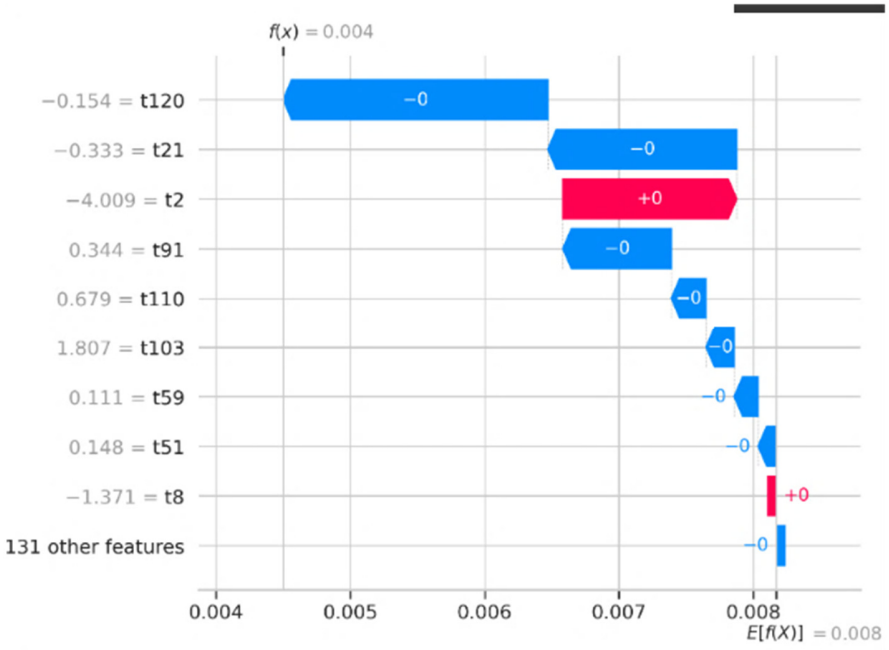

Fig. 4. SHAP waterfall plot of individual time-step contributions.

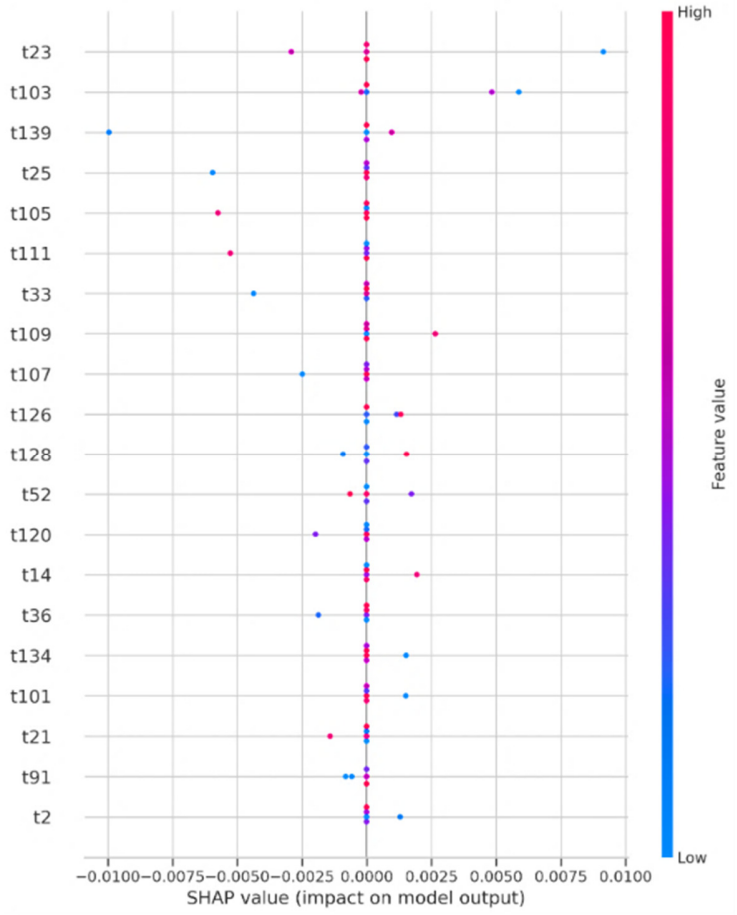

Fig. 5. SHAP summary plot of the distribution of time-step contributions across samples.

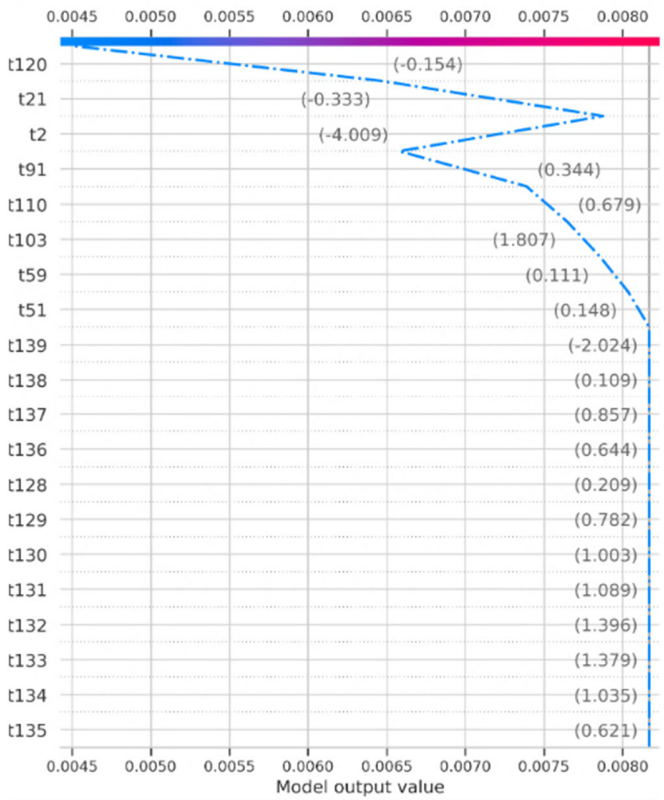

Fig. 6. SHAP decision plot of cumulative time-step impact for a single ECG sample.

B. Model Generalization and Scalability

To assess real-world applicability, the model was validated across heterogeneous ECG datasets with varying sampling rates, hardware, and patient demographics. The LSTM autoencoder showed minimal performance degradation on external datasets, maintaining low false positive and false negative rates even under noise and incomplete recordings. This robustness highlights its adaptability for deployment in hospitals, ambulatory care, and wearable devices. The fixed-window design and lightweight architecture support efficient scaling for continuous monitoring and large-scale screening.

C. Real-Time and Resource-Constrained Application

The proposed LSTM autoencoder is suitable for deployment in the following real-time and embedded environments:

- Low Latency Inference: Efficient processing of the proposed LSTM autoencoder can enable rapid anomaly alerts for timely intervention.
- Wearable Device Compatibility: A compact model and fixed input windows can ensure efficient operation on battery-powered devices.
- Remote Clinical Monitoring: Cloud integration of LSTM autoencoder will lead to secure transmission of anomaly scores and real-time clinician access.
- Energy Efficiency: Regularization, adaptive learning rates, and early stopping will reduce computational load and energy consumption.

IV. CONCLUSION AND FUTURE WORK

Cardiovascular diseases have generated a strong demand for reliable and automated Electrocardiogram (ECG) anomaly detection systems. Traditional machine learning approaches often rely on engineered features and struggle to capture the complex temporal dependencies present in ECG signals. Although deep learning models have improved detection performance, their lack of interpretability limits their adoption in clinical environments where transparent decision-making is crucial.

To address these challenges, this study proposed an explainable deep learning framework for ECG anomaly detection using a Long Short-Term Memory (LSTM) autoencoder integrated with SHapley Additive exPlanations (SHAP). The LSTM autoencoder learns the temporal patterns of normal cardiac rhythms and detects abnormalities through reconstruction error analysis, while SHAP provides interpretable explanations by highlighting the time-step contributions influencing anomaly predictions. Experimental evaluation on the ECG5000 dataset demonstrated strong performance, achieving high accuracy, precision, recall, and F1-score. The proposed approach outperformed traditional machine learning methods, such as Support Vector Machines (SVMs) and Logistic Regression (LR), by effectively capturing temporal ECG dynamics while maintaining model transparency. Additionally, SHAP-based visualizations provided insights into the decision-making process, enhancing clinical interpretability and trust.

Building upon this framework, future research should combine ECG with complementary data sources such as patient demographics, clinical history, and other physiological signals to enrich model context and improve diagnostic accuracy. In addition, adaptive tailored algorithms can be developed for individual patient baselines, addressing inter-patient variability and enhancing predictive sensitivity. Overall, the proposed framework combines effective anomaly detection with Explainable Artificial Intelligence (XAI), making it suitable for real-time ECG monitoring and clinical decision-support systems. Future work will focus on extending the framework to multiclass arrhythmia detection, integrating multimodal physiological data, and optimizing the model for deployment in wearable and edge-based healthcare applications.

DECLARATION OF COMPETING INTERESTS

The authors declare no competing interests.

ACKNOWLEDGEMENT

Not applicable in this paper.

DATA AVAILABILITY

The dataset used in this study is publicly available at: https://www.timeseriesclassification.com/description.php?Dataset=ECG5000.

REFERENCES

[1] A. Y. Hannun *et al.*, "Cardiologist-Level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network," *Nature Medicine*, vol. 25, no. 1, pp. 65–69, Jan. 2019, https://doi.org/10.1038/s41591-018-0268-3.

[2] C.-H. Hsieh, Y.-S. Li, B.-J. Hwang, and C.-H. Hsiao, "Detection of Atrial Fibrillation Using 1D Convolutional Neural Network," *Sensors*, vol. 20, no. 7, Apr. 2020, Art. no. 2136, https://doi.org/10.3390/s20072136.

[3] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long Short-Term Memory Networks for Anomaly Detection in Time Series," in *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges, Belgium, Apr. 2015, pp. 89–94.

[4] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions." arXiv, 2017, https://doi.org/10.48550/ARXIV.1705.07874.

[5] U. R. Acharya, H. Fujita, S. L. Oh, Y. Hagiwara, J. H. Tan, and M. Adam, "Application of Deep Convolutional Neural Network for Automated Detection of Myocardial Infarction Using ECG Signals," *Information Sciences*, vol. 415–416, pp. 190–198, Nov. 2017, https://doi.org/10.1016/j.ins.2017.06.027.

[6] G. B. Moody and R. G. Mark, "The Impact of the MIT-BIH Arrhythmia Database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, May 2001, https://doi.org/10.1109/51.932724.

[7] S. Osowski, L. T. Hoai, and T. Markiewicz, "Support Vector Machine-Based Expert System for Reliable Heartbeat Recognition," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 4, pp. 582–589, Apr. 2004, https://doi.org/10.1109/TBME.2004.824138.

[8] O. AlZoubi, N. AlAbabneh, I. Hmeidi, and M. Bani Yassein, "A Deep Learning System for the Diagnosis of Heart Problems from ECG Media Files," *International Journal on Communications Antenna and Propagation*, vol. 11, no. 5, Oct. 2021, Art. no. 363, https://doi.org/10.15866/irecap.v11i5.21132.

[9] Haibo He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, https://doi.org/10.1109/TKDE.2008.239.

[10] Y. Xia, N. Wulan, K. Wang, and H. Zhang, "Detecting Atrial Fibrillation by Deep Convolutional Neural Networks," *Computers in Biology and Medicine*, vol. 93, pp. 84–92, Feb. 2018, https://doi.org/10.1016/j.compbiomed.2017.12.007.

[11] 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*, San Francisco, CA, USA, Aug. 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

[12] C. K. N., Neelappa, M. T. Sreedevi, and R. Asha, "Hybrid Deep Learning Models for Accurate ECG Classification in Cardiovascular Disease Diagnosis," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 26683–26688, Oct. 2025, https://doi.org/10.48084/etasr.12209.

[13] S. M. Lundberg *et al.*, "From Local Explanations to Global Understanding with Explainable AI for Trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, Jan. 2020, https://doi.org/10.1038/s42256-019-0138-9.

[14] Y. Chen and E. Keogh, "Dataset: ECG5000." Time Series Classification, Mar. 2020, [Online]. Available: https://www.timeseriesclassification.com/description.php?Dataset=ECG5000.