Automated Arrhythmia Classification Using 2D Convolutional Neural Networks with Grad-CAM Explainability

A REPORT SUBMITTED IN

FULFILMENT OF THE REQUIREMENTS FOR INTERNSHIP

By

Dhananjay R

Rajiv Gandhi Institute of Technology, Kottayam

DURING

May 10, 2025 - August 10, 2025

UNDER THE GUIDANCE OF

Dr. Sreeja M U

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY KOTTAYAM

Abstract

Arrhythmia is a cardiac condition characterized by abnormalities in the rate, regularity, or origin of the heart's electrical impulses, resulting in irregular heartbeats. This work proposes a deep learning-based two-dimensional convolutional neural network (2D-CNN) for accurate classification of five different classes of arrhythmia. The dataset utilized in this study comprises heartbeat signals derived from two widely used repositories: the MIT-BIH Arrhythmia Database and the PTB Diagnostic ECG Database. Preprocessing techniques, including conversion of input images to grayscale to reduce input channel dimensionality, along with various data augmentation methods, were applied to improve generalization. The model's performance was evaluated using accuracy and F1 score, achieving a training accuracy of 99.22% and a validation accuracy of 99.10%, demonstrating high reliability in identifying common cardiac rhythms. Furthermore, explainability methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) and fidelity analysis were employed to interpret the model's decision-making process, providing insights into the specific regions of input data influencing predictions. These findings highlight the potential of the proposed 2D-CNN architecture as an effective and interpretable tool for automated arrhythmia classification.

Contents

1 Introduction

Heart rhythm disorders, known as arrhythmia, occur when the electrical signals that control the heartbeat become irregular. These conditions can range from harmless variations to serious health risks, making their timely detection very important. Electrocardiography (ECG) is the standard method used for recording and analyzing heart activity. While effective, manual interpretation of ECG signals can be slow and dependent on the expertise of medical professionals, which limits its scalability in large healthcare settings. Advances in artificial intelligence have introduced new opportunities for automating medical diagnosis.Deep learning, specifically, has shown great potential to analyze ECG signals. Convolutional neural networks (CNNs) are especially effective because they can automatically learn patterns in data without requiring hand-crafted features. This allows them to capture subtle differences in ECG waveforms that may indicate different types of arrhythmia.

The work presented in this report focuses on building a two-dimensional CNN model to classify five categories of arrhythmia. The study uses data compiled from two well-known ECG resources: the MIT-BIH Arrhythmia Database and the PTB Diagnostic Database. To improve the model's performance, preprocessing methods such as grayscale conversion and data augmentation were applied, which helped in reducing complexity and improving generalization. Model evaluation was carried out using accuracy and F1 score as performance metrics. The system achieved a training accuracy of 99.22% and a validation accuracy of 99.10%, which highlights its ability to classify arrhythmia types with a high degree of reliability. To make the predictions more transparent, explainability tools such as Grad-CAM and fidelity analysis were employed, allowing visualization of the regions that contributed most to the model's decisions. Overall, this study demonstrates the usefulness of deep learning in automating arrhythmia detection. By combining strong predictive performance with interpretability, the proposed approach can contribute to more efficient and trustworthy computer-aided diagnosis systems.

2 Literature Review

Electrocardiography (ECG) remains the standard diagnostic tool for detecting arrhythmias, as it provides a non-invasive means of recording the heart's electrical activity. Traditionally, cardiologists interpret ECG traces manually by assessing rhythm irregularities and waveform morphology. Although this approach is clinically reliable, it can be time-intensive and prone to variability among different observers, making it less suitable for large-scale or real-time applications . To address these limitations, researchers have increasingly focused on automated arrhythmia detection systems.

The initial wave of computer-assisted arrhythmia detection primarily used traditional machine learning algorithms such as Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Decision Trees, and Random Forests. Sraitih et al., 2021. These systems often relied on handcrafted features extracted from ECG signals, including temporal parameters (e.g., QRS width, PR interval) and frequency-domain descriptors. While these models provided useful results, their effectiveness was restricted by the quality of manually engineered features and their sensitivity to noise and inter-patient variability. Ahsan and Siddique, 2022.

The rise of deep learning significantly advanced ECG classification by eliminating the need for manual feature design. One-dimensional convolutional neural networks (1D-CNNs) have been widely applied to sequential ECG signals, consistently outperforming earlier machine learning models. Acharya et al., 2017 Recurrent architectures such as RNNs and Long Short-Term Memory (LSTM) networks have also been utilized to capture temporal dependencies in ECG sequences. Yildirim, 2018.

In parallel, two-dimensional convolutional neural networks (2D-CNNs) have gained popularity by transforming ECG signals into image-based formats, such as heartbeat segments or spectrograms. Ullah et al., 2020. These models exploit spatial information and have shown strong generalization capabilities across datasets. A consistent trend in recent literature indicates that CNN-based approaches surpass conventional methods in terms of accuracy and robustness, particularly when applied to large and diverse patient data. Rajpurkar et al., 2017.

Research in this domain largely depends on standardized databases. The MIT-BIH Arrhythmia Database is widely adopted for benchmarking as it provides annotated heartbeat signals collected from clinical practice. Moody and Mark, 2001. Similarly, the PTB Diagnostic ECG Database includes recordings from both healthy individuals and patients with various cardiac conditions, enabling broader diagnostic applications. Wagner et al., 2020. Despite their importance, these datasets present challenges such as class imbalance, inter-patient variation, and noise, which complicate model development and training.

With deep learning models achieving high predictive accuracy, the demand for interpretability has grown significantly. Black-box predictions often limit acceptance among healthcare professionals, especially in sensitive areas like cardiology. To address the black-box nature of deep learning models, explainability techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) have been employed. This method highlights the regions of the input that most influence the model's predictions, providing insights into the decision-making process.Selvaraju et al., 2017. These strategies not only improve transparency but also enhance clinician trust in automated systems.

This study proposes a two-dimensional CNN architecture for the classification of five arrhythmia categories. Preprocessing steps such as grayscale conversion and data augmentation were applied to reduce data imbalance and improve generalization. Importantly, the study goes beyond accuracy-focused evaluation by incorporating Grad-CAM and fidelity analysis to interpret model decisions. By addressing both performance and interpretability, this work aims to advance the development of clinically trustworthy arrhythmia detection systems.

3 Methodology

3.1 Dataset

The dataset used in this study combines heartbeat signals from two widely recognized sources: PhysioNet's MIT-BIH Arrhythmia Database and the PTB Diagnostic ECG Database. These datasets provide electrocardiogram (ECG) recordings representing both normal heartbeats and those affected by arrhythmias or myocardial infarction.

To prepare the data, the ECG signals were preprocessed and segmented so that each segment corresponds to a single heartbeat.Based on the annotations provided in the MIT-BIH and PTB Diagnostic ECG databases, the beats were grouped into five categories: Normal/Non-ectopic (N), Myocardial Infarction (M), Ventricular ectopic (V), Supraventricular ectopic (S), and Unclassifiable (Q).

For model development, the dataset was divided into training and testing subsets, with 80% of the beats used for training and 20% reserved for testing. The detailed distribution of beats across the five categories is provided in Table 1.

Table 1: Distribution of Arrhythmia Types in the Dataset

Arrhythmia Type	**Total Samples**
Myocardial infarction (M)	10,506
Normal/Non-ectopic beats (N)	94,635
Supraventricular ectopic beats (S)	2,779
Ventricular ectopic beats (V)	7,236
Unclassifiable beats (Q)	8,039

3.2 Image data preprocessing

All ECG beat images were converted into 128×128 grayscale format, since color information is irrelevant for distinguishing arrhythmia types in this study. Converting to grayscale reduces the number of input channels, thereby lowering computational complexity and simplifying analysis. To further standardize the data, images were rescaled to the range [0, 1], which facilitates efficient training of neural networks. The processed 2D ECG beat images were then directly fed into the deep learning model without additional preprocessing.

To improve generalization and mitigate overfitting, data augmentation techniques were applied. These included random horizontal and vertical shifts of up to 10% of the image size, as well as random brightness adjustments within the range 0.8–1.2. In addition, class imbalance handling was carried out by applying class weights and reducing the number of normal beats (N) from approximately 75,000 to around 9,000, ensuring that minority classes received greater emphasis during training.

3.3 Model Architecture

In this study, we designed a Convolutional Neural Network (CNN) to perform arrhythmia classification. The complete framework is illustrated in Figure 1.

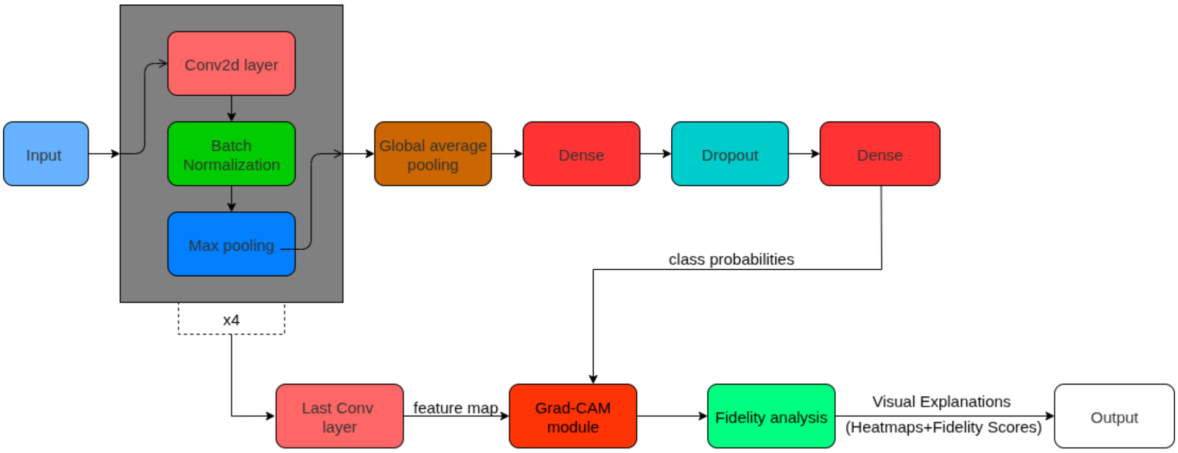


Figure 1: System architecture of the proposed CNN model.

The network architecture consists of four convolutional layers, where the number of filters increases at each stage. Every convolutional layer is followed by batch normalization to improve training stability. For feature reduction, the model applies three max-pooling operations and a global average pooling layer. To strengthen generalization, the design also includes two fully connected layers and a dropout layer.

The activation strategy is as follows: ReLU was employed in all convolutional layers and the first dense layer to accelerate training and capture non-linear relationships while avoiding vanishing gradients. The final layer uses a Softmax function, converting raw outputs into probability distributions across the classes, which makes it appropriate for multi-class classification problems.

For training, the learning rate was set to 0.001. If validation loss remained unchanged for three epochs, the rate was decreased by a factor of 0.5. Through experimentation, a batch size of 64 was chosen as the most effective configuration.

3.4 Explainability and Fidelity Analysis

Deep learning models, though highly accurate, are often criticized as “black-box” systems due to their lack of interpretability. In medical applications such as arrhythmia classification, it is crucial to ensure that the model’s decisions are not only accurate but also understandable to clinicians. Explainable AI (XAI) techniques address this need by providing insights into the internal reasoning of the model. One widely used method is Gradient-weighted Class Activation Mapping (Grad-CAM), which highlights the specific regions of an input image that contribute most strongly to the model’s prediction. By doing so, Grad-CAM helps verify whether the model is focusing on clinically meaningful features (such as P waves, QRS complexes, or T waves in ECG signals), thereby enhancing trust, transparency, and reliability in its predictions.

3.4.1 Grad-CAM Visualization

To enhance the interpretability of the proposed CNN model, Gradient-weighted Class Activation Mapping (Grad-CAM) was implemented. Grad-CAM allows visualization of the regions in the input ECG images that contribute most strongly to the model’s classification decisions. In this work, the last convolutional layer was identified as the target layer for generating activation maps, since it preserves spatial information relevant for classification. For a given input ECG image, a forward pass was performed to obtain the class score corresponding to the predicted arrhythmia class. The gradients of this class score with respect to the feature maps of the selected layer were then computed using backpropagation. These gradients were spatially averaged to derive importance weights for each feature map, which quantify the contribution of each channel to the model's decision. A weighted sum of the feature maps was calculated, followed by a ReLU activation to retain only the positively influential regions. The resulting Grad-CAM heatmap was normalized and upsampled to match the size of the input ECG image. Finally, the heatmap was superimposed onto the original grayscale ECG image, providing a visual explanation of which waveform regions, such as the P wave, QRS complex, or T wave, the CNN model relied upon for classification.

3.4.2 Fidelity Analysis

While Grad-CAM provides valuable visual insights, a quantitative measure is necessary to assess the reliability of these explanations. To address this, fidelity analysis was performed, which evaluates whether the regions highlighted in the Grad-CAM heatmap are genuinely important for the model's predictions. In this work, binary masks were generated from the Grad-CAM heatmaps by thresholding the most activated regions, representing the "important" areas of the ECG images. Two modified versions of the input images were then created for evaluation. The preserved mask retained only the highlighted regions while masking out the rest, whereas the deleted mask removed the highlighted regions, keeping the remaining parts of the image intact. The CNN model was re-evaluated on both versions, and changes in classification confidence were measured. A significant drop in prediction confidence when the highlighted regions were removed indicated high fidelity. Finally, the fidelity score was quantified by comparing the model's prediction probabilities between the preserved and deleted inputs. A higher fidelity score confirmed that the Grad-CAM explanations were consistent and trustworthy, demonstrating that the model's decisions were indeed influenced by the regions identified in the heatmaps.

4 Results

4.1 Model Performance

The proposed 2D Convolutional Neural Network (CNN) was trained and validated on the prepared dataset consisting of arrhythmia images derived from the MIT-BIH Arrhythmia Database and the PTB Diagnostic ECG Database. After data preprocessing, augmentation, and class balancing, the model achieved a training accuracy of 99.22% and a validation accuracy of 99.10%. The close correspondence between these values indicates that the model generalizes well and is not significantly overfitting. To further assess performance, precision, recall, and F1-score were computed for each of the five arrhythmia classes (N, S, V, Q, and M). The overall weighted F1-score was found to be high, confirming that the classifier performs consistently across classes.

The results are summarized in Table 2.

Table 2: Classification Metrics for the CNN Model

Metric	**Macro Average**	**Weighted Average**
Precision	0.81	0.96
Recall	0.95	0.93
F1-Score	0.86	0.94

The confusion matrix shown in figure 2 highlights that most samples were correctly classified, with only a few instances of misclassification between classes exhibiting similar waveform patterns.The model was configured to train for 100 epochs, but due to the use of the early stopping mechanism, training halted after 25 epochs once no further improvements were observed. Throughout training, the best-performing model was automatically saved. The learning rate, which was initially set at 1.0×10^{-3}, was progressively reduced to 1.0×10^{-4} when the validation loss failed to improve across several consecutive epochs.

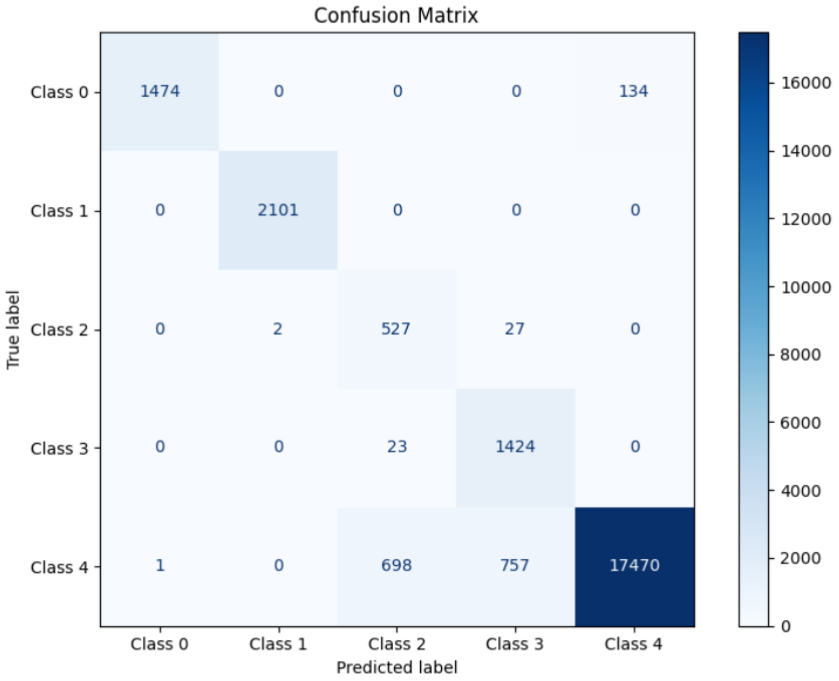
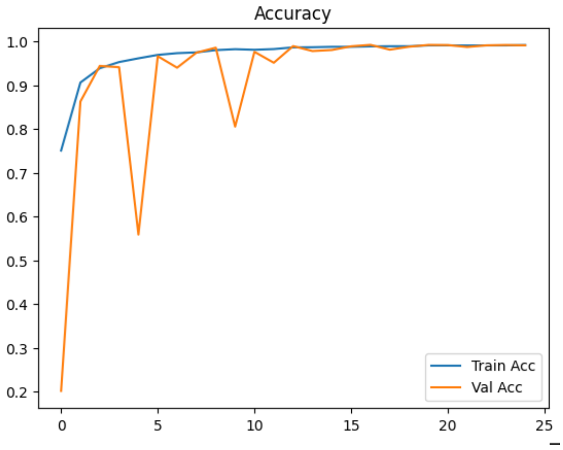


Figure 2: Confusion Matrix for the CNN Model

The plots depicting the model's accuracy and loss over the training epochs are shown in Figures 3 and 4, respectively.



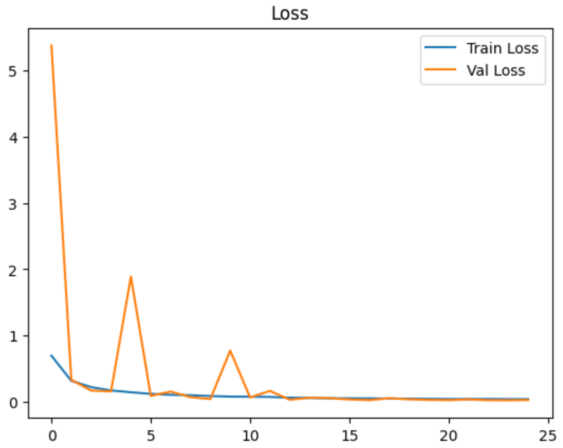


Figure 3: Accuracy vs Number of Epochs

Figure 4: Loss vs Number of Epochs

4.2 Explainability Analysis using Grad-CAM

Convolutional Neural Networks (CNNs) often function as black-box models, making it difficult to interpret how predictions are derived from image data. To address this, Gradient-weighted Class Activation Mapping (Grad-CAM) was introduced, a widely used method for visualizing and debugging CNNs. This technique leverages the gradients of the target class flowing into the final convolutional layer to generate a heat map that highlights the most influential regions of an image in the model's decision-making. By computing a weighted combination of the feature maps, Grad-CAM identifies the key spatial areas that contribute to each prediction, allowing the resulting heat map to be superimposed on the original image for intuitive interpretation.

Representative Grad-CAM heatmaps for ECG images are illustrated in Figure 5

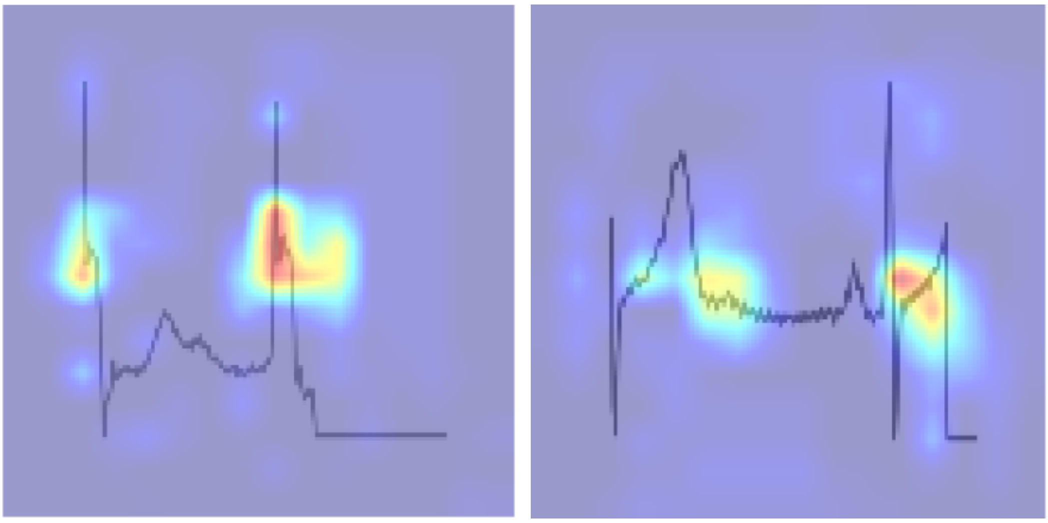


Figure 5: Representative Grad-CAM heatmaps for ECG images. (Left) Q-class, (Right) N-class.

4.3 Fidelity Evaluation

To quantitatively assess the reliability of Grad-CAM, fidelity scores were computed across three different threshold levels for representative images from each arrhythmia class. The results demonstrated that fidelity values remained relatively stable across thresholds, with only minor variations. This consistency implies that the highlighted regions in the Grad-CAM outputs are robust indicators of the model's predictions.

A comparative plot of fidelity values across classes in figure 6 illustrates this stability and further supports the interpretability of the approach.

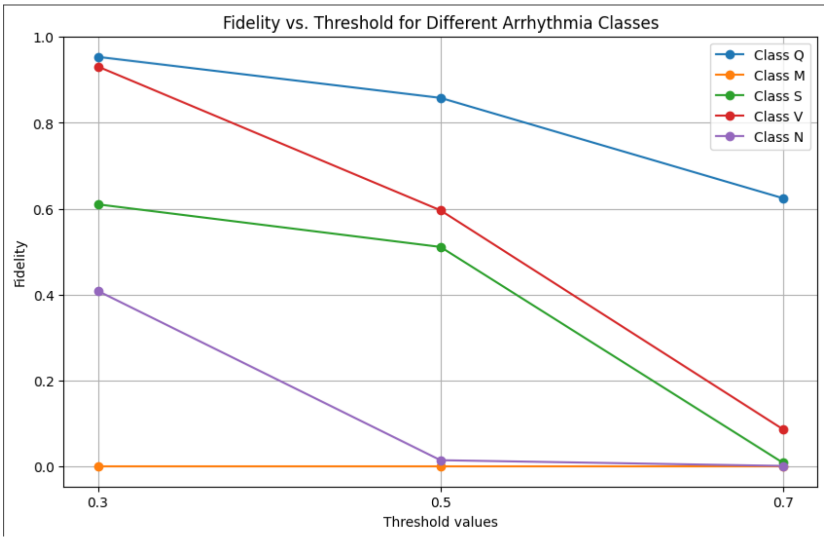


Figure 6: Fidelity values measured at different threshold levels across sample ECG classes.

In summary, the CNN demonstrated excellent classification performance, and the integration of Grad-CAM with fidelity analysis provided valuable interpretability. These findings suggest that deep learning-based models can support automated arrhythmia diagnosis, though validation on larger and more diverse datasets remains an essential future step.

5 Conclusions

This project successfully implemented a two-dimensional Convolutional Neural Network (2D-CNN) for ECG arrhythmia classification, achieving high performance across multiple metrics. Data preprocessing, including grayscale conversion, normalization, augmentation, and class balancing, enabled robust learning and reduced overfitting, resulting in a training accuracy of 99.22% and validation accuracy of 99.10%.

Explainability was incorporated using Gradient-weighted Class Activation Mapping (Grad-CAM), which highlighted clinically relevant waveform regions such as P waves, QRS complexes, and T waves. Fidelity analysis further confirmed that the highlighted regions were genuinely influential in the model's predictions, ensuring that the visual explanations were reliable and consistent with the model's decision-making process.

Evaluation through the confusion matrix and class-wise metrics showed strong performance for normal beats, while some arrhythmia classes with similar waveforms remained challenging to differentiate. This highlights areas for future improvement, such as enhanced feature extraction, integration of temporal information, or the use of hybrid architectures combining CNNs with sequential models.

Overall, the study demonstrates that 2D-CNNs, when combined with explainability methods, can provide both accurate and interpretable arrhythmia classification. The results support the potential of such models to aid clinicians in automated ECG analysis, reduce manual workload, and improve diagnostic efficiency. Future work may focus on incorporating larger, more diverse datasets and evaluating the system in real-world clinical settings to enhance generalization and practical applicability.

References

Acharya, U. R., Fujita, H., Lih, O. S., Hagiwara, Y., Tan, J. H., & Adam, M. (2017). Automated detection of arrhythmias using different intervals of tachycardia ecg segments with convolutional neural network. *Information sciences*, *405*, 81–90.

Ahsan, M. M., & Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, *128*, 102289.

Moody, G. B., & Mark, R. G. (2001). The impact of the mit-bih arrhythmia database. *IEEE engineering in medicine and biology magazine*, *20*(3), 45–50.

Rajpurkar, P., Hannun, A. Y., Haghpanahi, M., Bourn, C., & Ng, A. Y. (2017). Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836*.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*, 618–626.

Sraitih, M., Jabrane, Y., & Hajjam El Hassani, A. (2021). An automated system for ecg arrhythmia detection using machine learning techniques. *Journal of Clinical Medicine*, *10*(22), 5450.

Ullah, A., Anwar, S. M., Bilal, M., & Mehmood, R. M. (2020). Classification of arrhythmia by using deep learning with 2-d ecg spectral image representation. *Remote sensing*, *12*(10), 1685.

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W., & Schaeffter, T. (2020). Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, *7*(1), 1–15.

Yildirim, Ö. (2018). A novel wavelet sequence based on deep bidirectional lstm network model for ecg signal classification. *Computers in biology and medicine*, *96*, 189–202.