ECG Heartbeat Classification: A Comparative Study of Machine Learning and Deep Learning Approaches for Arrhythmia Detection

Ishita Pahuja†, Diya Tyagi†
Indira Gandhi Delhi Technical University for Women, New Delhi, India
ishita073btcseai23@igdtuw.ac.in, diya055btcseai23@igdtuw.ac.in
† These authors contributed equally to this work.

***Abstract*—Electrocardiogram (ECG) signals provide useful information about the functioning of the heart and help identify abnormal heart rhythms. While these signals are commonly used in cardiac monitoring, analysing them manually can be challenging, especially when large amounts of data are involved. This project explores how machine learning and deep learning can be used to classify ECG heartbeats and distinguish between different heartbeat patterns.**

The work uses the MIT-BIH heartbeat dataset, which contains five categories of heartbeats: Normal, Supraventricular, Ventricular, Fusion and Unknown. Different traditional machine learning algorithms and deep learning models are explored. These include CNN, GRU, hybrid CNN-RNN and attention-based architectures. Since the dataset contains an unequal distribution of heartbeat classes, techniques such as SMOTE, data augmentation and class weighting are also considered. Grad-CAM is used alongside the deep learning models to highlight the parts of an ECG signal that influence their predictions.

Through these approaches, the project examines the differences between various classification models, particularly in their ability to recognise less frequent heartbeat categories. It also looks at how visual explanations can help in understanding the predictions made by deep learning models. The overall aim is to explore different approaches to ECG heartbeat classification while considering both model performance and interpretability.

***Index Terms*—ECG Classification, Arrhythmia Detection, Machine Learning, Deep Learning, CNN, GRU, Class Imbalance, Explainable AI.**

I. Introduction

The heart generates electrical signals that reflect its electrical activity. These signals can be recorded using an electrocardiogram (ECG), which enables healthcare professionals to monitor cardiac activity and identify irregular heart rhythms. Different heartbeat categories exhibit distinct waveform patterns, and analysing these patterns can help identify potential cardiac abnormalities. However, manual examination of ECG recordings is time-consuming, particularly when processing large volumes of data. Consequently, automated ECG analysis has become an important area of research.

Machine learning and deep learning techniques have increasingly been applied to automate ECG signal analysis. Traditional machine learning algorithms, such as Random Forest, Logistic Regression, and Support Vector Machines (SVM), rely on extracted features to classify heartbeats. In contrast, deep learning models can automatically learn relevant patterns from input signals. For instance, Convolutional Neural Networks (CNNs) can identify morphological patterns in ECG waveforms, while Gated Recurrent Units (GRUs) can capture sequential dependencies within signals. Hybrid CNN-RNN architectures combine these capabilities, while attention mechanisms enable models to assign greater importance to waveform regions that are relevant to their predictions.

A major challenge in ECG classification is class imbalance, as different heartbeat categories are represented by unequal numbers of samples. Consequently, models may achieve high performance on majority classes while struggling to identify less frequent categories. This limitation is particularly important in medical applications, where the misclassification of abnormal heartbeats may affect the reliability of automated systems. Techniques such as Synthetic Minority Over-sampling Technique (SMOTE), data augmentation, and class weighting can help address this issue by improving the representation of minority classes during model training.

In this study, we investigate ECG heartbeat classification using the MIT-BIH Arrhythmia heartbeat dataset, which comprises five categories: Normal, Supraventricular, Ventricular, Fusion, and Unknown. The study examines multiple classical machine learning and deep learning approaches, including Random Forest, Logistic Regression, SVM, CNN, GRU, hybrid CNN-RNN, and attention-based architectures. Different strategies for handling class imbalance are also evaluated. Furthermore, Grad-CAM is employed to visualise the temporal regions of ECG waveforms that contribute to the predictions of deep learning models.

The primary objective is to compare the classification performance of these approaches across different heartbeat categories, with particular attention to the recognition of underrepresented classes. The study also examines the effectiveness of imbalance-handling techniques and uses Grad-CAM to provide qualitative insights into the regions of ECG signals influencing model predictions.

II. Literature Review

De Chazal et al. [1] investigated automatic heartbeat classification using morphological and temporal features extracted

from ECG signals. Their approach considered characteristics such as heartbeat shape and intervals between consecutive beats to classify heartbeats into five categories according to the AAMI standard. Their findings demonstrated the importance of waveform morphology and heartbeat timing in distinguishing different cardiac patterns.

Luz et al. [2] reviewed existing approaches to ECG-based heartbeat classification and arrhythmia detection. Their review covered the principal stages of ECG analysis, including signal preprocessing, heartbeat segmentation, feature extraction, and classification. They also discussed methodological challenges and factors influencing the performance of automated classification systems.

Kiranyaz et al. [3] explored one-dimensional Convolutional Neural Networks (CNNs) for ECG classification. Their approach enabled networks to learn relevant features directly from ECG signals, reducing dependence on manually engineered descriptors. They also investigated patient-specific classification, demonstrating the potential of CNN-based approaches for automated heartbeat analysis.

Acharya et al. [4] developed a deep CNN architecture for classifying different heartbeat categories. Their approach focused on learning discriminative patterns from ECG waveforms to distinguish between different cardiac conditions. The study highlighted the potential of deep learning to automate feature extraction and improve heartbeat recognition.

Hannun et al. [5] proposed a deep neural network for detecting and classifying cardiac arrhythmias from ECG recordings. They evaluated the model's predictions against interpretations provided by cardiologists. Their findings demonstrated the potential of deep learning approaches for automated ECG analysis and arrhythmia detection.

Chawla et al. [6] introduced the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance in datasets. The method generates synthetic samples for underrepresented classes using information from neighbouring data points. This approach is relevant to ECG classification, where normal heartbeats typically outnumber several abnormal heartbeat categories.

Murat et al. [7] reviewed deep learning techniques for ECG-based heartbeat detection and classification. Their study examined different neural network architectures and their applications in identifying abnormal cardiac rhythms. It also highlighted challenges associated with signal preprocessing, class imbalance, and model selection.

Selvaraju et al. [8] introduced Gradient-weighted Class Activation Mapping (Grad-CAM), an explainability technique for interpreting CNN predictions. The method uses gradients to identify input regions that contribute to a model's output. It can also be adapted to ECG signals to highlight waveform segments that influence classification decisions.

Ayano et al. [9] reviewed interpretable machine learning techniques for ECG-based heart disease classification. Their work examined approaches for improving the interpretability of model predictions, emphasising their importance in medical applications. The study highlighted the need for classification systems that combine predictive performance with interpretability.

III. METHODOLOGY

A. Dataset and preprocessing

The experiments were conducted using the MIT-BIH Arrhythmia heartbeat dataset in its commonly distributed CSV format, comprising 109,446 heartbeat records. The original training and testing CSV files were concatenated and shuffled before the experiments. Each record contains 187 numerical waveform observations followed by a class label, resulting in 188 columns. The original dataset has a sampling frequency of 125 Hz.

The five heartbeat categories are represented by the labels N (Normal), S (Supraventricular Ectopic), V (Ventricular Ectopic), F (Fusion), and Q (Unknown). Their corresponding descriptions are presented in Table I.

TABLE I
HEARTBEAT CATEGORIES AND THEIR CORRESPONDING LABELS

Label	Heartbeat Category
N	Normal beat
S	Supraventricular ectopic beat
V	Ventricular ectopic beat
F	Fusion beat
Q	Unknown beat

Exploratory data analysis included class-frequency inspection, representative waveform visualisation, descriptive statistics, correlation analysis on a 5,000-record subset, autocorrelation analysis, first-order signal differences, and frequency-domain analysis using the Fast Fourier Transform (FFT). Principal Component Analysis (PCA) and t-distributed Stochastic Neighbour Embedding (t-SNE) were used to visualise lower-dimensional representations. K-Means clustering (k = 2–10) was also explored using the elbow criterion and internal clustering indices. These analyses were descriptive and were not treated as independent evidence of clinical class separability.

An exploratory outlier detection procedure calculated Z-scores within each heartbeat across its 187 samples. A record was flagged when its maximum absolute Z-score exceeded 9. The implemented cleaning procedure removed flagged records only from class N, reducing the combined dataset from 109,446 to 109,211 records. Since this class-specific procedure does not constitute a validated ECG artefact-removal method, it was treated as an exploratory ablation.

For deep learning experiments, StandardScaler was fitted exclusively on the training subset and subsequently applied to the validation and test subsets.

B. Feature representation and classical machine learning

For classical machine learning, 17 features were extracted from each heartbeat: mean, standard deviation, skewness, kurtosis, median, R-peak value, minimum amplitude, peak-to-peak range, mean slope, maximum slope, zero-crossing count, signal energy, dominant frequency, spectral energy in the 0–10 Hz, 10–30 Hz, and 30–60 Hz bands, and spectral entropy.

Mutual-information-based SelectKBest was used to retain 10 features: mean, standard deviation, skewness, kurtosis, median, signal energy, the three frequency-band energy features, and spectral entropy.

Three input representations were evaluated: the original 187-point waveform (raw), the 10 selected features (selected), and their concatenation (combined). The classical machine learning algorithms included Logistic Regression, Decision Tree, Random Forest, AdaBoost, and K-Nearest Neighbors (KNN).

The classical experiments used stratified 80:20 training-test splits. StandardScaler was incorporated into the pipelines for raw and combined inputs, whereas the selected-feature pipeline used the extracted features without additional scaling. SMOTE was integrated into the training pipeline for experiments involving class imbalance.

Additional Random Forest experiments investigated the effects of SMOTE, PCA, and the exploratory outlier-removal procedure. PCA was evaluated as a dimensionality-reduction technique, including a representation that retained 95% of the variance. These experiments examined alternative preprocessing and imbalance-handling strategies rather than introducing additional model architectures.

C. Deep-learning architectures and training

Deep learning experiments used 187 waveform samples as a one-dimensional input sequence. The dataset was divided using stratified sampling into 70% (76,611 records) for training, 10% (10,945 records) for validation, and 20% (21,890 records) for testing, using `random_state = 42`.

TABLE II
ARCHITECTURES AND THEIR IMPLEMENTED STRUCTURES

Architecture	Implemented Structure
1D-CNN	Four Conv1D blocks with 64, 256, 512 and 512 filters (kernel sizes 3, 9, 11 and 25), batch normalization, max pooling and dropout; global average pooling; 256-unit dense layer; five-class softmax output.
GRU	GRU layers with 128 and 64 units, followed by a 514-unit ReLU dense layer, dropout and five-class softmax output.
Hybrid CNN–RNN	Three waveform segments processed by convolutional branches. Their representations are concatenated and passed to a bidirectional GRU (32 units), dense layers, dropout and softmax output.
Attention Model	A full-signal convolutional/bidirectional-GRU branch with temporal attention, combined with three local convolutional branches representing selected waveform regions, followed by dense layers and softmax output.

D. Evaluation and model interpretation

Model performance was evaluated using accuracy, class-wise precision, recall, and F1-score, along with macro-averaged F1-score and recall. Confusion matrices were also generated to examine classification performance across the five heartbeat categories.

Macro-F1 was selected as the primary evaluation metric because it assigns equal importance to all five classes, irrespective of their individual frequencies. The results for classes S and F were reported separately due to their substantial underrepresentation in the dataset.

For a given class c, precision was calculated as:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}$$

Recall was calculated as:

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}$$

The F1-score was calculated as the harmonic mean of precision and recall. Macro-F1 was obtained by computing the unweighted average of the five class-wise F1-scores.

Grad-CAM was applied to selected predictions from the deep learning models to visualise the temporal regions contributing to their classification decisions. The implementation used the tf-keras-vis library, with a linearised model output and class-specific scoring applied to selected convolutional layers.

The notebook reports visualisations for two sample indices (423 and 412). These visualisations were used as qualitative illustrations of model activations rather than as statistically validated measures of explanation faithfulness.

E. Methodological qualification

The original training and testing CSV files were concatenated before generating new training, validation, and test subsets. Consequently, the resulting test subset does not represent the untouched official test set.

Additionally, the mutual-information-based feature selector was fitted before the classical train-test split, introducing a possibility of feature-selection leakage. The notebook does not establish whether the data were partitioned on a patient-wise basis.

Therefore, the reported results should be interpreted within the context of the implemented experimental setup rather than as definitive estimates of patient-independent or clinical generalisation. Future experiments should preserve the official test set, perform feature selection exclusively within the training folds, and implement patient-wise separation wherever patient identifiers are available.

IV. RESULTS AND DISCUSSIONS

A. Dataset Composition and Evaluation Protocol

The experiments were performed using the MIT-BIH Arrhythmia data set in its Kaggle CSV format. After merging the training and test files provided, there were 109,446 heartbeat recordings in the merged data set. Each recording consisted of 187 waveform samples and a single class label. There were five different types of classes: normal (N), supraventricular ectopic (S), ventricular ectopic (V), fusion (F) and unknown/other (Q).

TABLE III
CLASS DISTRIBUTION IN THE DEEP-LEARNING TEST SET.

Class	Support	Percentage (%)
N	18,118	82.77
S	556	2.54
V	1,447	6.61
F	161	0.74
Q	1,608	7.35
Total	21,890	100.00

TABLE IV
LEADING CLASSICAL MACHINE-LEARNING CONFIGURATIONS RANKED BY MACRO-F1.

Classifier	Representation	Macro-F1
Random Forest	Combined	0.9155
Random Forest	Raw	0.9134
K-Nearest Neighbours	Raw	0.8148
Decision Tree	Combined	0.7890
K-Nearest Neighbours	Combined	0.7888
Decision Tree	Raw	0.7881
Random Forest	Selected	0.7698

In the deep learning experiments, the merged dataset was split into three parts according to a stratified 70:10:20 ratio for training, validation and testing purposes. As a result, the test set had 21,890 records. Table III provides the class distribution of the test set. For the classical machine learning experiments, the same stratified 80:20 split was used to create a test set. This means that both classical and deep learning results need to be interpreted within each method's specific evaluation protocol rather than as a directly comparable head-to-head comparison on a common test set.

The distribution has considerable class imbalance. F and S account for only a small percentage of the test instances, and accuracy alone is not a good metric to assess the model's performance.

Therefore, we use macro-averaged precision, recall and F1-score as well as class-specific recall for S and F in our comparison. assess the model's performance.

Therefore, we use macro-averaged precision, recall and F1-score as well as class-specific recall for S and F in our comparison.

B. Classical Machine-Learning Results

The five classical classifiers were evaluated using three different feature representations: the original 187-dimensional waveform (Raw), the 10-dimensional feature-selected representation (Selected), and a combined representation containing the waveform and engineered features (Combined). The Random Forest classifier using the Combined representation yielded the highest macro-F1 score of 0.9155 for the five classical classifiers. Random Forest with Raw features followed closely, with a macro-F1 score of 0.9134. The leading configurations are summarized in Table IV.

The results of the best classical model can be found in the classification report of the best classical model. The results show a high F1-score for the majority N class (F1-score of 0.99) as well as for the Q class (F1-score of 0.98). For the S, V and F classes the F1-scores are 0.85, 0.95 and 0.81 respectively. Even though the overall accuracy of the model is 0.98, the lower support for S and F means that their class-specific recall is of particular importance.

As part of the analysis of the different preprocessing steps, we also conducted an ablation study with Random Forest on the notebook. First, we applied SMOTE and found that the reported macro-F1 increased from 0.874 to 0.915. Next, we applied PCA and found that the macro-F1 decreased to 0.855. Finally, we removed the outliers and found that the macro-F1 increased slightly from 0.874 to 0.876. These are the results of the experiments conducted on the notebook and should not be confused with the estimates of the classical models, as the feature-selection was fitted before the classical train-test split was conducted. This means that there is a potential information-leakage pathway which might lead to overly optimistic classical-model estimates.

C. Deep-Learning Model Comparison

Deep-learning experiments with 1D CNN variants, GRU-based model, CNN–RNN hybrid, attention-focused architecture and frozen Hugging Face ECG encoder with XGBoost were compared as shown in Table V using the exact same values as reported in the final model-comparison table in the notebook.

The Hybrid CNN–RNN had the highest macro-F1 score (0.9064), followed by the attention-focused model (0.9056) and the class-weighted CNN (0.9047). The class-weighted CNN achieved the highest macro recall (0.9460), The attention-focused model also achieved a high macro recall of 0.9363. The encoder-plus-XGBoost model had the lowest macro-F1 score among the five models (0.8637). The encoder-plus-XGBoost configuration produced the lowest macro-F1 score among these five configurations (0.8637).

These results confirm that the model with the highest macro-F1 score is not necessarily the model with the highest macro recall. The choice of model therefore depends on the relative importance assigned to precision and missed minority-class examples.

D. Performance on Minority Arrhythmia Classes

We report separately the recall values for S and F arrhythmia classes as they are under-represented in the dataset. The class-weighted CNN model achieved the highest S recall (0.9065) and F recall (0.8882) among the five final deep-learning configurations. CNN achieved the highest S recall (0.9065) and F recall (0.8882) among the five final deep-learning configurations. The attention-focused model followed with S recall of 0.8705 and F recall of 0.8696. The encoder-plus-XGBoost model obtained lower recall for both classes.

TABLE V
DEEP-LEARNING AND ENCODER-BASED MODEL COMPARISON ON THE COMMON DEEP-LEARNING TEST PARTITION.

Model	Macro-F1	Macro Recall	Macro Precision	Accuracy
Hybrid CNN–RNN (class-weighted)	0.9064	0.9151	0.8985	0.98
Attention-focused	0.9056	0.9363	0.8793	0.98
CNN (class-weighted)	0.9047	0.9460	0.8730	0.98
GRU hybrid (class-weighted)	0.8912	0.9233	0.8640	0.98
Hugging Face encoder + XGBoost	0.8637	0.8416	0.8891	0.97

TABLE VI
MINORITY-CLASS PERFORMANCE OF DEEP-LEARNING MODELS

Model	S-Rec.	S-Prec.	F-Rec.
CNN (Weighted)	0.9065	0.7790	0.8882
GRU Hybrid	0.8471	0.6999	0.8447
Hybrid CNN–RNN	0.8058	0.8043	0.8385
Attention	0.8705	0.7344	0.8696
HF Encoder + XGBoost	0.6709	0.8180	0.7205

The class-weighted CNN has the highest recall for the minority classes in this comparison. S precision for the class-weighted CNN is lower than that of the Hybrid CNN–RNN, indicating a trade-off between increasing the number of detected minority-class examples and increasing the number of false-positive predictions. The attention-focused model provides a more balanced alternative, with high recall for both minority classes. These are classification results on the available test partition and should not be used as clinical safety evidence.

E. Explainability Analysis

We also applied Grad-CAM to some examples of the one-dimensional convolutional models and visualized the temporal regions that contribute to a prediction of the model for a single test instance. In the following we show two examples for test instances 423 and 412. Note that this is a qualitative evaluation and no population-level analysis of explanation faithfulness is performed. We do not report an aggregate measure for explanation faithfulness over a large sample of instances nor do we perform a systematic comparison.

Thus, no claims about the explanation reliability can be made.

F. Limitations and Interpretation

Several limitations exist with regard to the reported results. A direct comparison of the classical pipeline's results with the results of the deep-learning pipeline cannot be made because the two were trained and tested on different train-test partitions. feature selection for the classical pipeline was performed before the train-test split, introducing potential The notebook does not demonstrate a patient-wise partition and therefore the test performance reported does not represent generalization to unseen patients. patient-wise partition; consequently, the reported test performance should not be assumed to represent generalization to unseen patients. Fourth, the CNN experiment summaries contain differences between some narrative validation scores and the callback-derived values, while the final model-comparison table reports a separate set of test metrics. The exact values in the final comparison table are retained in Table V; the differing validation summaries should not be conflated with test-set performance.

Finally, Grad-CAM analysis is presented on a few selected examples. The results presented are based on a small number of selected examples. Further work is required to establish patient-wise evaluation, leakage-free feature selection within the cross-validation folds, repeated runs with multiple random seeds, and confidence intervals. Furthermore, a systematic explanation evaluation is required before any clinical interpretation is attempted.

G. Summary of Findings

Among the classical configurations, Random Forest with combined features reported the highest macro-F1 score (0.9155). Within the deep-learning comparison, Hybrid CNN–RNN had the highest macro-F1 (0.9064), while the class-weighted CNN had the highest macro recall (0.9460) and the strongest recall for the S and F classes (0.9065 and 0.8882, respectively). These results indicate that class weighting can improve sensitivity to minority classes; however, the evaluation limitations above need to be addressed in order to evaluate the generalization and clinical applicability of the models.

V. CONCLUSION

This study compared the performance of several classical machine-learning algorithms and a number of deep-learning architectures for ECG heartbeat classification into five categories of heartbeats. The best performing model among the classical machine-learning algorithms was the Random Forest model using a combination of the raw ECG waveform and engineered features, which achieved a macro-F1 score of 0.9155. The best performing deep-learning architecture was the Hybrid CNN–RNN model, which achieved a macro-F1 score of 0.9064. The class-weighted CNN achieved the highest recall for the minority S and F classes. This study highlights the effectiveness of combining feature engineering with ensemble learning for ECG classification, as well as the need to handle the class-imbalance problem in ECG classification. Furthermore, the experiments conducted in this study show that overall accuracy is not a good measure of the performance of a classification model for ECG classification, as it can be heavily influenced by the performance of the model on the most frequent class, while the performance of the model on the under-represented classes can be very poor. Further work

is required to investigate the generalizability and reliability of the models developed in this study. In particular, we plan to conduct patient-wise evaluation, leakage-free feature selection, experiments with multiple random seeds, as well as more systematic explainability analysis.

VI. FUTURE WORK

Although the models discussed in this work cover different approaches to ECG classification, there is still room to explore how well they perform on a wider range of ECG recordings. The models can be tested on data collected from different patients to see if they are able to recognise heartbeat patterns they have not encountered during training. It would also be interesting to check how they handle recordings affected by noise, since real ECG signals are not always clean.

One aspect that requires more attention is the classification of less frequent heartbeat types. Since these categories have fewer samples in the dataset, models may find it harder to recognise them correctly. The models could also be compared using patient-independent testing to get a clearer picture of how well they generalise.

The work can also be taken towards real-time ECG analysis, where heartbeat patterns are classified as the signals are recorded. This would require looking at how quickly the models make predictions and whether they can maintain good classification performance while processing continuous ECG data.

REFERENCES

[1] P. de Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic Classification of Heartbeats Using ECG Morphology and Heartbeat Interval Features," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 7, pp. 1196–1206, 2004. doi: 10.1109/TBME.2004.827359.

[2] E. J. da S. Luz, W. R. Schwartz, G. Camara-Chavez, and D. Menotti, "ECG-Based Heartbeat Classification for Arrhythmia Detection: A Survey," *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 144–164, 2016. doi: 10.1016/j.cmpb.2015.12.008.

[3] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 664–675, 2016. doi: 10.1109/TBME.2015.2468589.

[4] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, M. Adam, A. Gertych, and R. S. Tan, "A Deep Convolutional Neural Network Model to Classify Heartbeats," *Computers in Biology and Medicine*, vol. 89, pp. 389–396, 2017. doi: 10.1016/j.compbiomed.2017.08.022.

[5] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-Level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network," *Nature Medicine*, vol. 25, no. 1, pp. 65–69, 2019. doi: 10.1038/s41591-018-0268-3.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. doi: 10.1613/jair.953.

[7] F. Murat, O. Yildirim, M. Talo, U. B. Baloglu, Y. Demir, and U. R. Acharya, "Application of Deep Learning Techniques for Heartbeats Detection Using ECG Signals—Analysis and Review," *Computers in Biology and Medicine*, vol. 120, Art. no. 103726, 2020. doi: 10.1016/j.compbiomed.2020.103726.

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.

[9] Y. M. Ayano, F. Schwenker, B. D. Dufera, and T. G. Debelee, "Interpretable Machine Learning Techniques in ECG-Based Heart Disease Classification: A Systematic Review," *Diagnostics*, vol. 13, no. 1, Art. no. 111, 2023. doi: 10.3390/diagnostics13010111.