Comparative Analysis of Deep Learning and Traditional Machine Learning Models for Arrhythmia Classification using ECG Signals

Deepak Murali
Heritage High School

Arrhythmias, a form of cardiovascular disease, are a major contributor to the high global mortality rate. Early detection of arrhythmias through electrocardiogram (ECG) analysis can significantly improve patient outcomes. This study investigates the application of var- ious machine learning (ML) and deep learning models for the classification of arrhythmias using ECG signals from the MIT-BIH Arrhythmia Dataset. The models evaluated include Random Forest, Support Vector Machines (SVM), Logistic Regression, Multilayer Perceptron (MLP), and Convolutional Neural Networks (CNN). Additionally, feature selection techniques, such as the Fourier Transform, were applied to enhance the performance of the ML models. Among the models tested, the CNN achieved the highest accuracy (89.29%), F1 score (85.69%), and AUC (87.98%), demonstrating its superior ability in accurately detecting arrhythmias. In contrast, traditional ML models, including Random Forest and SVM, showed moderate performance with lower accuracy and discriminatory power. The study highlights the potential of CNN-based architectures for automated ECG analysis and emphasizes the importance of integrating explainable AI techniques to increase the transparency and clinical adoption of deep learning models. Future research could focus on larger, more diverse datasets and the use of Recurrent Neural Networks (RNNs) for longer ECG recordings to improve classification performance further.

Introduction

Cardiovascular diseases are the leading cause of mortality worldwide, with over 16% of all deaths being caused by heart disease (2020). Among various heart diseases, arrhythmias pose significant risks. Arrhythmias are irregular heartbeats and can often be symptoms of underlying heart conditions or other health problems. Untreated arrhythmias cause complications like heart failure, stroke, or sudden cardiac arrest (Y. P. Sai et al. 2020, pp. 1-6). Detecting arrhythmias early can significantly improve patient outcomes and reduce healthcare costs. Given the benefits of early detection, exploring the use of machine learning (ML) models for arrhythmia classification becomes vital, as these models can be used to automatically and accurately detect arrhythmia. This paper aims to use deep learning methods to classify arrhythmia using electrocardiogram (ECG) signals from the MIT-BIH Arrhythmia Dataset. The dataset includes ECG recordings that cover a variety of arrhythmia, allowing machine-learning algorithms to be trained to identify patterns in ECG signals that indicate arrhythmia.

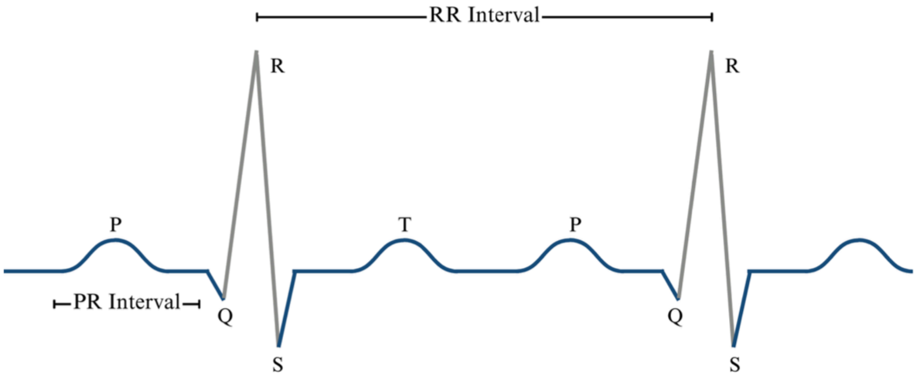


Figure 1. An ECG waveform highlighting the important intervals, with the QRS complex colored shaded in gray.

Significant advancements in machine learning and deep learning approaches for arrhythmia detection have already been made. Despite this, gaps remain in scalability, model adaptability, and performance consistency across diverse datasets. For instance, Tsipouras et al. (2005, pp. 237-250) achieved high accuracy using RR intervals for arrhythmia classification. The RR interval, which represents the time between successive R-peaks in an ECG signal (as shown in Figure 1), is a crucial feature in heart rate variability analysis and is often used to detect irregularities in cardiac rhythms. While effective in identifying arrhythmias related to heart rate fluctuations, this method struggled to generalize to other types, such as atrial fibrillation (AFib)—a condition characterized by rapid and irregular electrical impulses that disrupt the heart’s normal rhythm. Unlike arrhythmias primarily reflected in RR interval variations. Similarly, the Optimum-Path Forest (OPF) model introduced by Luz et al. (2013, pp. 3561–3573) showed promising F1 scores, but its complexity and scalability challenges made it less effective on larger datasets. In another study, Shimpi et al. (2017, pp. 603–607) employed a Bag of Visual Words (BoVW) approach, which relied on pre-extracted features. However, this method’s reliance on specific domain-based features limited its adaptability to different datasets and ECG signal variations, underscoring the need for more flexible and generalizable models.

This study moves beyond these limitations by leveraging a deep learning architecture, specifically Convolutional Neural Networks (CNNs), which can automatically extract features from raw ECG signals. Unlike methods that rely heavily on pre-defined features like RR intervals or specific waveform characteristics, CNNs can learn relevant patterns directly from the data, making the model more flexible and capable of identifying a wider range of arrhythmias. Additionally, the end-to-end nature of the CNN architecture ensures that the model can handle both temporal and spatial patterns in the ECG signals, providing a more comprehensive analysis compared to traditional machine learning models. This allows the model to be scalable and adaptable to larger datasets with complex or noisy ECG signals, addressing the issues of generalization and performance in real-world settings.

2 Related Works

2.1 Cardiac Arrhythmias Diagnosis

Traditionally, doctors diagnose cardiac arrhythmias with a comprehensive review of the patient's clinical history and physical examination, which provides crucial information about the presence of any signs and symptoms (Levy, 1991). This assessment directs them in selecting the appropriate test. A primary tool used for diagnosis is the electrocardiogram (ECG).

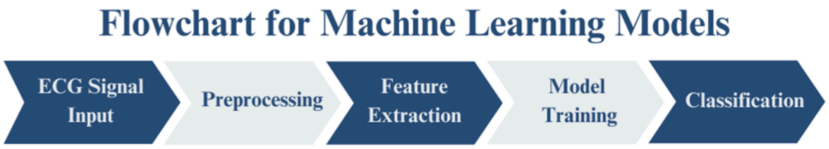


Figure 2. Flowchart of machine learning techniques used for ECG signal classification.

ECGs record the electrical activity of the heart to identify the type and severity of the arrhythmia (Levy, 1991). Interpreting an ECG reading involves analyzing various components, including the P wave, PR interval, QRS complex, ST segment, T wave, and U wave (Ward, 2015, pp. 473–475). Each of these intervals corresponds to specific cardiac events. For instance, the P wave reflects atrial depolarization, the PR interval represents conduction through the atrioventricular node, the QRS complex indicates ventricular depolarization, and the T wave shows ventricular repolarization (Tadros, 2017). When all the components occur in their normal intervals, it is referred to as a sinus rhythm. Abnormalities in a sinus rhythm, such as irregularities in the timing or structure of these components—specifically the P wave, PR interval, and QRS complex—indicate the presence of arrhythmia. Analyzing these irregularities can help determine the type of arrhythmia and develop effective treatment strategies.

2.2 Machine Learning Methods

Machine learning has emerged as a powerful tool for various applications, especially medical diagnostics. For example, a study by Maity and Das explored the use of machine-learning techniques to diagnose Alzheimer's disease (2021, pp. 1393-1398). Using data from the National Alzheimer's Coordinating Center, they built a Bayesian model to compute and determine the probability of an Alzheimer's diagnosis given the patient's family history, cognitive test, and other risk factors. The model achieved an accuracy of nearly 80% in predicting diagnoses.

Machine learning techniques have also been used to analyze ECG signals to classify arrhythmias due to their success in pattern recognition and other medical diagnoses. Traditional methods to develop such algorithms involved manually extracting features based on domain knowledge and clinical understanding of ECG waveforms (See Figure 2 for all the steps). These features typically included time-domain metrics like the RR interval, QRS complex duration, and PR interval (as illustrated in Figure 1). A study done by Tsipouras et al. focused on leveraging time-domain features, particularly the RR interval, for arrhythmia classification using the MIT-BIH Arrhythmia Database (Tsipouras et al. 2005, pp. 237–250) The RR interval represents the time between R peaks in an ECG waveform and is used to assess heart rhythm regularity. In addition to RR intervals, they explored other relevant features to differentiate between different types of arrhythmias. This approach achieved a high accuracy of 95% in arrhythmic beat classification and emphasized the significance of extracting certain features using clinical knowledge to automate arrhythmia detection. However, this method is limited to detecting episodes specifically related to the RR interval and can- not detect other arrhythmias like atrial fibrillation. The accuracy also decreases in larger datasets due to the increased noise and arrhythmias the method cannot detect.

In contrast to traditional feature extraction methods, a study by Luz et al. introduced Optimum-Path Forest (OPF), a graph-based classification technique developed using the MIT-BIH dataset to address issues with previous models (2013, pp. 3561–3573) . OPF uses a graph approach where the nodes represent samples and the edges represent the proximity between these samples in the feature space. This methodology allows OPF to efficiently capture the underlying structure of the data. Compared to other well-known classifiers like Support Vector Machines (SVM) and

Multilayer Perceptron, the OPF classifier had better F1 scores, a metric that evaluates a model's ability to correctly classify positive and negative instances while considering both errors and correct predictions, but lower accuracy scores ranging from 80% to 90% depending on the features extracted.

OPF may be more robust and able to detect abnormalities in a variety of ECG signals, but its complexity makes it difficult to scale. These scalability challenges cause OPF to be less accurate when dealing with larger datasets, and it increases the computational cost significantly, essentially leading to OPF losing its advantage.

Building on advancements in feature extraction and machine learning classification, a study done by Shimpi et al. introduces another feature extraction and machine learning method for classifying ECG data into different types of arrhythmia (2017, pp. 603–607). In this study, the authors used the UCI Machine Learning Repository dataset, which included ECG data from 472 patients recorded with 279 attributes each. They used principal component analysis (PCA) on the dataset to reduce it to only 150 predictions and preserve around 99% of the variance. They also use the Bag of Visual Words (BoVW) model to classify. This model is designed to segment ECG signals and extract key domain-based features from each segment, like the QRS complex and RR interval. These features are then clustered using K-means, assigning each feature to one of the defined clusters. These clusters have values assigned to them, and histograms can be generated to represent how many of each value there are in a segment. Machine learning classifiers like SVM and Random Forests are then used to categorize these histograms. SVM showed the highest classification accuracy of 91.2%. However, the study's reliance on pre-extracted features from the dataset can make it difficult for it to adapt to different datasets and feature extraction methods. Direct feature extraction from raw signals can improve the model's flexibility.

In a more recent study, Kumari et al. used Support Vector Machines, a supervised machine learning algorithm, to achieve over 95% accuracy in arrhythmia classification (2021, pp. 1393–1398). To extract features, the study incorporated Discrete Wavelet Transform (DWT), which breaks down the ECG signal into useful components at different frequencies. DWT works by applying low-pass and high-pass filters to the signal. The low-pass filters capture the overall shape of the ECG signal, which can help to see some of the broader trends in the data. On the other hand, the high-pass filters capture fine details, such as sudden spikes or noise, providing a detailed view of the signal's complex features. The thorough feature extraction makes it easier for the SVM classifier to differentiate between various arrhythmia types. The model utilized datasets from MIT-BIH, MIT-BH Sinus, and BIDMC for training and validation.

The application of machine learning techniques in medical diagnostics, particularly for arrhythmia classification using ECG signals, shows the significant advancements made in integrating AI-based technology into healthcare. Studies have shown the accuracy of various models, such as Random Forest, Optimum-Path Forest, and SVMs, in accurately classifying heart arrhythmias. Feature extraction has significantly evolved due to advanced algorithms like DWT and PCA, yet challenges in scalability and adaptability to different datasets highlight the need for more research in this field.

2.3 Deep Learning Methods

Another notable development in this domain would be the introduction of deep learning methods. Despite the improved feature extraction and model selection in machine learning, researchers decided to switch to deep learning models due to their ability to automatically learn from the data without explicit programming and find useful features on their own. Deep learning has also been highly successful in other domains, such as image processing and Natural Language Processing (NLP). In image processing, techniques such as Convolutional Neural Networks (CNNs) have been used to achieve success in image classification and object detection. Models like AlexNet, VGGnet, and ResNet have effectively recognized objects and patterns in images with high precision (Jiao et al. 2019, pp. 172 231–172 263). Deep learning has also transformed the field of NLP by advancing how machines understand human language. Deep Learning models have excelled in language modeling, parsing, and semantic processing. Models like BERT and GPT-3 have demonstrated how deep learning can excel in text generation along with language processing, leading to significant progress in generative AI. These successes across diverse fields showcase the capabilities of deep learning and how it is applied to solve complex problems.

In a study by Ouelli et al., the use of Multilayer Perceptron (MLP) was employed for arrhythmia classification (2014, pp. 402–406.). The model they built used finite impulse response for noise reduction in ECG signals and multivariate autoregressive modeling to extract relevant features from two-lead ECG signals. FIR filters reduce noise by averaging the signal over a fixed number of points, keeping important parts of the ECG. MV AR modeling analyzes how the two ECG leads to change over time to identify patterns in the heart's activity. The MLP model was trained and evaluated using the MIT-BIH database, Creighton University Ventricular Tachyarrhythmia Database, and MIT-BIH Supraventricular Arrhythmia Database. This proposed method achieves an overall classification accuracy of 99.7%.

A study done by Rajkumar et al. introduced a deep-learning approach using multichannel CNN for ECG classification (2019, pp. 365–369). Multichannel CNNs can handle more input channels than regular CNN, making them more suitable for medical imaging or signal processing. The raw ECG signals from the MIT-BIH arrhythmia database were directly fed into the CNN, which processed the features. This approach used a stochastic gradient descent (SGD) algorithm to minimize the loss function and had an accuracy of 93.6%.

2.4 Challenges

Classifying arrhythmias from ECG signals poses several challenges that impact the reliability of machine learning and deep learning models. These include signal noise and dataset imbalance, both of which can affect model performance. However, this study employs preprocessing and data augmentation techniques to mitigate these issues.

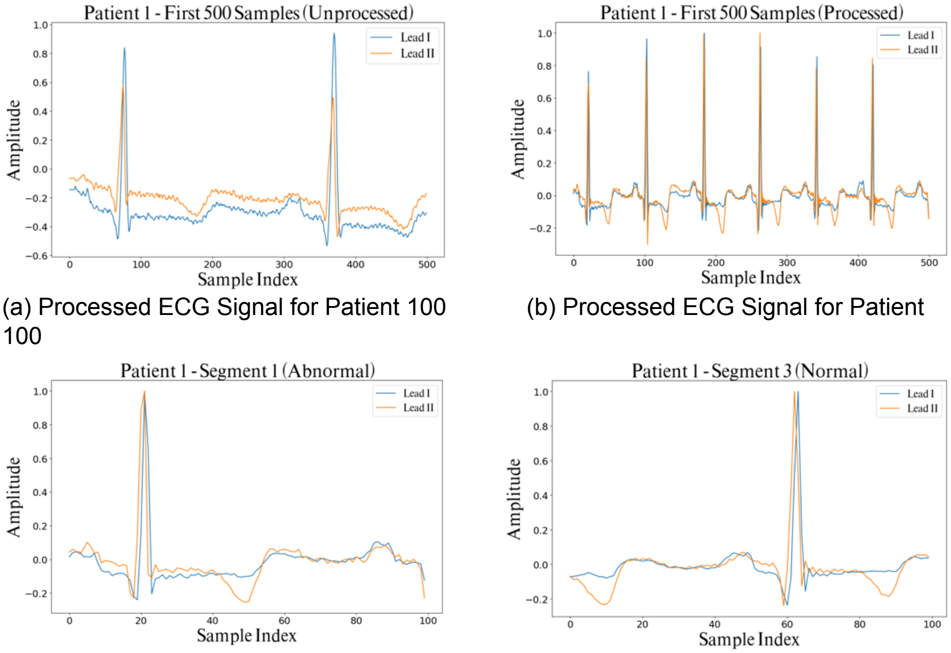
ECG signals often have a low signal-to-noise ratio, which can obscure important features and reduce classification accuracy. Some common noise sources include baseline wander, power-line interference, and muscle artifacts. Respiration, body movements, poor electrode contact, and skin-electrode impedance contribute to these disturbances. Baseline wander, which has a frequency spectrum ranging between 0.05 Hz and 1 Hz, can affect the QRS complex amplitude, making it appear higher or lower than its true value. Power-line interference, caused by electrical sources operating at frequencies of 50 to 60 Hz, can distort the P-wave. Muscle artifacts, also known as electromyographic noise, occur due to electrical activity from muscle contractions and overlap with the ECG frequency range, making it difficult to isolate the true heart signal. To mitigate these challenges, a bandpass filter is applied to remove noise outside the typical heart rate frequency range, improving signal clarity and model reliability.

Another challenge is dataset imbalance, which is prevalent in the MIT-BIH Arrhythmia Database. The dataset contains a disproportionate number of normal heartbeats compared to arrhythmic beats, which can cause machine learning models to favor normal classes while failing to detect rarer arrhythmias. To address this issue, the Synthetic Minority Over- sampling Technique (SMOTE) is applied to generate synthetic samples for the minority class, improving the model's ability to recognize less frequent patterns. While this helps balance class distributions, it does not introduce additional variability in the dataset.

3 Method

3.1 Dataset

Researchers have explored various machine learning methods to automate arrhythmia detection, often using the MIT-BIH Arrhythmia Database to evaluate and train their models. This database provides annotated ECG signals from 48 patients, recorded over 30-minute durations on two channels, sampled at a frequency of 360 Hz, and includes the age and gender of each patient (Sahoo et al. 2020, pp. 185–194) (see Figure 3(a)). The annotations in the database identify different types of arrhythmias and normal beats, allowing researchers to accurately la- bel and classify the ECG data. By using this dataset, machine learning algorithms can be trained to classify arrhythmia (Apandi et al. 2018, pp. 1–5). The ECG recordings analyzed in this study include Lead I and Lead II, which provide different perspectives of the heart's electrical activity. Lead I measures the voltage between the left and right arms, showing electrical activity moving sideways across the heart. Lead II measures the voltage between the right arm and left leg, following the heart's natural electrical pathway. Because of this, Lead II is often used to monitor heart rhythm and detect irregularities like arrhythmias (Sampson et al. 2015, pp. 588–594).



(a) Processed ECG Signal for Patient 100 100

(b) Processed ECG Signal for Patient

(a) Segmented Normal ECG Signal
(d) Segmented abnormal ECG Signal

Figure 3. Raw and processed ECG signals for Patient 100, along with segmented normal and abnormal ECG waveforms. Lead I and Lead II capture the heart's electrical activity from different angles.

3.2 Preprocessing

First, a bandpass filter is applied to the ECG signals with a frequency range between 0.1 and 30.0 Hz. This filter helps remove noise and artifacts outside the typical range of heart rate frequencies, improving signal clarity and reducing the impact of irrelevant frequency components (Mihov, 2020, pp. 1–4). After filtering, the signals are down sampled to a lower frequency of 100 Hz, as shown in Figure 3(b) (Kwon, 2018, pp. 198–206), which helps reduce computational complexity and ensures that the data remains in a manageable format for further analysis. The dataset is then segmented into fixed-length windows of 1 second, facilitating more manageable and consistent data input for machine learning models (see Figure 3(c) or 3(d)). Each segmented ECG window is then classified as either normal or abnormal based on the annotations given to each sample. For example, in Figure 3(d), the annotations marked multiple samples in that segment to be abnormal due to irregular P-QRS-T wave patterns that require further clinical evaluation. Normal segments exhibit a consistent P-QRS-T wave pattern as shown in Figure 3(c).

The overall dataset consists of 5,092 normal segments and 2,876 abnormal segments, which presents a significant class imbalance. This imbalance can cause the model to be biased toward predicting the majority class (normal segments), reducing its ability to accurately detect abnormal patterns. To address this issue, the Synthetic Minority Over- sampling Technique (SMOTE) is employed. SMOTE works by generating synthetic samples for the minority class. It does this by interpolating between a sample from the minority class and one of its nearest neighbors, creating a new, synthetic data point that is a linear combination of the two (Glagus et al. 2013, pp. 1–16). This approach balances the data by mitigating model bias toward the majority class, enhancing the model's ability to recognize less frequent patterns, such as abnormal heartbeats, and improving overall detection performance.

3.3 Benchmark Models

3.3.1 Support Vector Machines

Support Vector Machines (SVMs) are a class of supervised machine learning methods developed for binary classification tasks and later expanded for other uses. SVMs are highly powerful when it comes to pattern learning due to their capability to handle high- dimensional data and generalizations. They are used in various applications such as text detection, image classification, and bioinformatics (Nasiri, 2009, pp. 187–192). SVMs try to find the optimal hyperplane that separates two classes in feature space with the largest margin, with the margin being the distance between the hyperplane and the closest data points from each class. Maximizing the margin helps achieve better generalization (Mammone, 2009, pp. 283–289). The structure of an SVM, highlighting the separating hyperplane and margin, is shown in Figure 4(a).

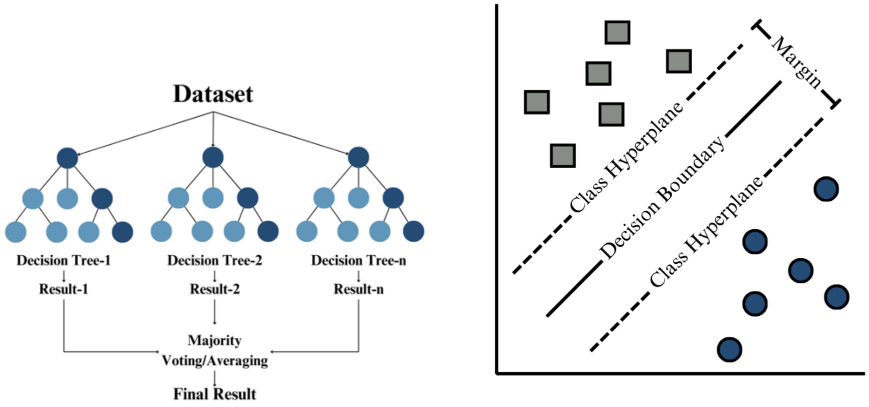


Figure 4. Comparison of Support Vector Machines (SVM) and Random Forest Classifier.

3.3.2 Random Forest Classifier

Random forests are a machine learning method that uses multiple decision trees to improve classification accuracy. Each tree in the forest is trained on a random subset of the data, and the final prediction is made based on the majority decision of these individual trees. This approach reduces the impact of noise and improves robustness compared to single decision trees or other methods like AdaBoost. Additionally, Random Forests internally monitor their own performance and error rates, allowing them to assess how the model's accuracy changes with the number of features. This method can be effectively used for both classification and regression tasks, making it powerful in many data analysis tasks (Breiman, 2001, pp. 5–32). The structure of a Random Forest, showing the ensemble of decision trees, is illustrated in Figure 5(b).

3.3.3 Multilayer Perceptron (MLP)

MLP is a simple neural network that consists of multiple layers of nodes. MLPs typically have an input layer, one or more hidden layers, and an output layer (as illustrated in Figure 5(a), left). MLPs are designed to model non-linear relationships in data and are effective in pattern recognition tasks. They're trained using backpropagation, a supervised learning technique that adjusts the weights of the connections to reduce the error between predicted and actual outputs. Backpropagation allows MLPs to automatically learn useful features from the data (Ramchoun, 2016).

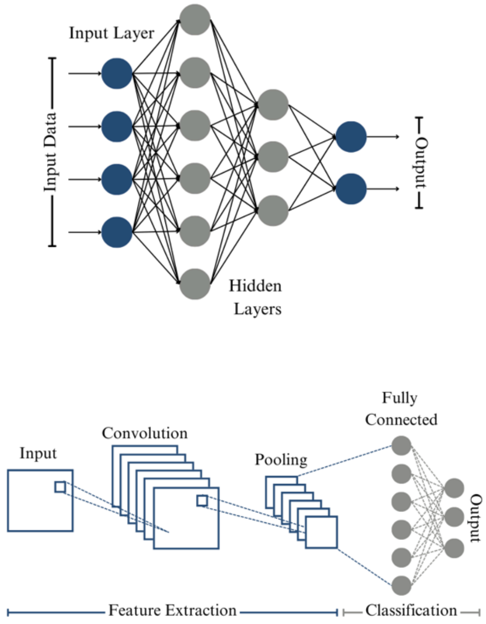


Figure 5. Comparison of Multilayer Perceptron (MLP) and Convolutional Neural Network (CNN).

3.3.4 Convolutional Neural Networks (CNNs)

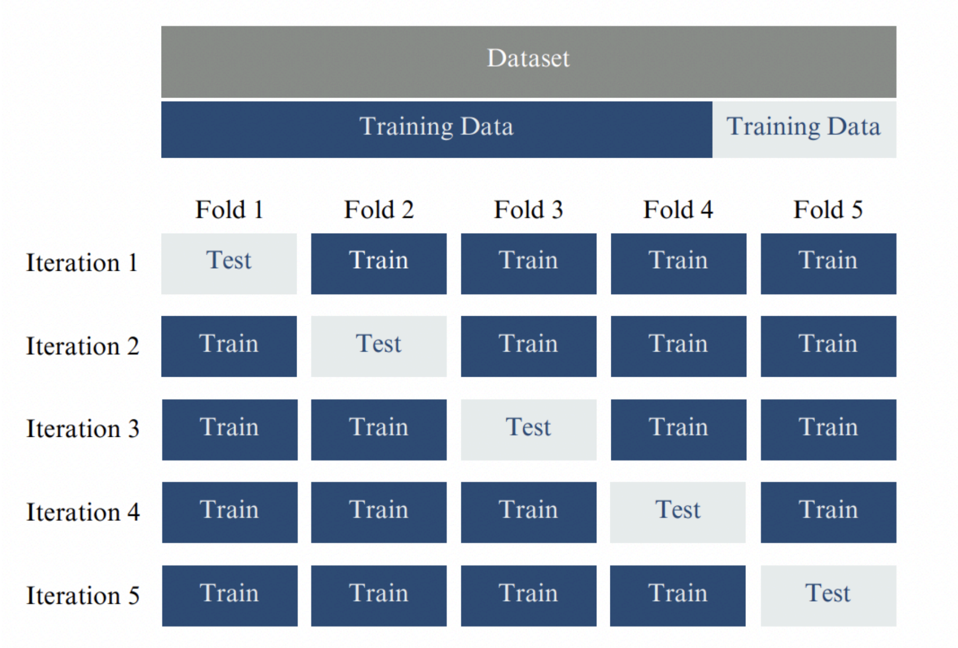
CNNs are another subclass of neural networks that excel in processing grid-like data, making them effective for a wide range of applications like image, speech, and natural language processing. Compared to MLPs, CNNs have a more complex structure, with convolutional layers, pooling layers, and specialized layers like batch normalization and dropout layers (basic structure shown in Figure 5(b), right). Convolutional layers extract features from the input using various filters. Then, a pooling layer is used to reduce the spatial dimensions of the features to decrease computational cost. Batch normalization layers normalize the output, and dropout layers are used to prevent overfitting by randomly removing neurons during the training stage. This layered structure enables CNNs to process data effectively, leading to superior performance in various applications (Krichen, 2023, p. 151).

3.4 Experiment Setting

After the preprocessing steps, the ECG signals were segmented into fixed-length segments, with each segment labeled as normal or abnormal based on annotations. This segmentation then returned a dataset of labeled signal segments. For model evaluation, 10-fold cross- validation was employed to ensure robust performance assessment (see Figure 6 for a visual representation of the process). The dataset was split into an 80:20

ratio for training and validation. The training set was further resampled using SMOTE to address class imbalance, ensuring a balanced dataset before model training. The validation set was used to monitor model performance and tune hyperparameters. Additionally, no separate holdout test set was used; instead, performance metrics were averaged across all folds to provide a reliable estimate of model generalization.

Three machine learning models—SVM, Logistic Regression, and Random Forest—were trained and evaluated using this cross-validation approach. Each model was selected for its unique strengths in ECG classification: SVM handles complex patterns and non-linear relationships, Logistic Regression serves as a simple and interpretable baseline, and Random Forest is robust to noise while identifying important features. This combination ensures a well-balanced evaluation of ECG classification performance. During each fold, SMOTE was applied to the training data whenever class imbalance was detected. To standardize feature values, StandardScaler was fitted on the training data to standardize feature values. To optimize model performance, hyperparameter tuning was performed using GridSearchCV, a systematic approach that evaluates different hyperparameter combinations through exhaustive search. The models were trained on the processed training data and predictions were generated for the test data in each fold. Performance metrics, including accuracy, precision, recall, and F1-score, were computed for each fold and then averaged to assess overall model performance.



	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test

Figure 6. An Illustration of a 5-fold Cross-Validation

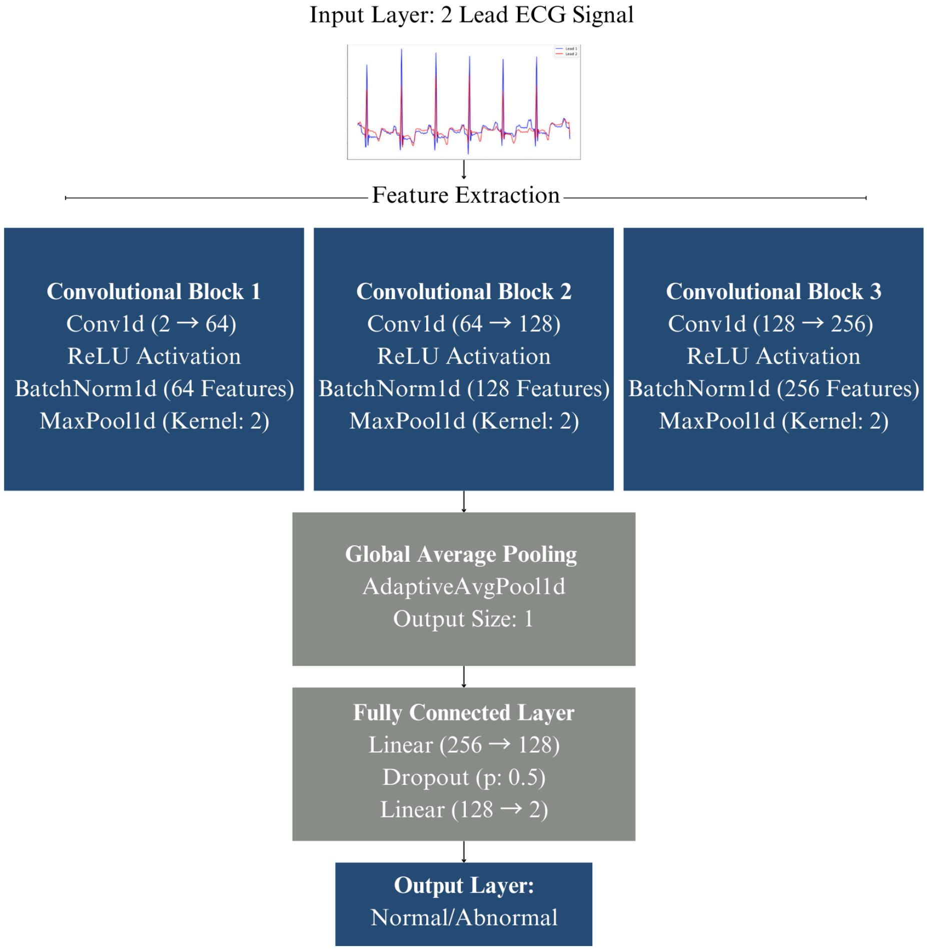


Figure 7. Structure of the CNN model used for ECG signal classification.

4 Implementation

A CNN model was developed and optimized through a structured training process to classify ECG signals effectively. Hyperparameter tuning was conducted empirically through iterative experimentation, adjusting key parameters based on validation performance. The number of convolutional layers, kernel size, dropout rate, and learning rate were tuned to achieve an optimal balance between model complexity and generalization. Initial experiments were performed with kernel sizes ranging from 3 to 9, with a size of 7 selected based on validation accuracy. Similarly, dropout rates between 0.3 and 0.6 were tested, with 0.5 providing the best trade-off between performance and overfitting. Learning rate tuning was performed using a ReduceLROnPlateau scheduler, which dynamically adjusted the rate by a factor of 0.5 if validation loss stagnated for five consecutive epochs. These hyperparameter choices were determined by monitoring performance on the validation set over multiple runs. The model was trained using a mini-batch gradient descent approach, where batches of 32 samples were fed into a network to update their weights iteratively. The Adam optimizer was used with a learning rate of 0.001 and a weight decay of 1e-4, providing a balance between convergence and regularization to prevent overfitting. The learning rate was adjusted using a ReduceLROnPlateau scheduler, which decreased the learning rate by a factor of 0.5 if the validation loss did not improve for five consecutive epochs. To further prevent overfitting, early stopping was implemented, halting training if validation accuracy did not improve for 15 consecutive epochs. The model was trained for 100 epochs on a GPU, ensuring efficient processing.

The training process was implemented using PyTorch and sklearn libraries. The architecture consisted of 3 convolutional blocks, each comprising a 1D convolutional layer, batch normalization, and max pooling. Batch normalization is a technique used to normalize activations and improve accuracy while speeding up the training process (Bjork, 2018). Max pooling is a method used in convolutional neural networks to reduce the size of feature maps by picking the highest value in a specific area. This helps keep important features while making the model more efficient and less sensitive to small changes in the input data (Murray et al. 2014, pp. 2473–2480) (See Figure 7 For full workflow).

The SMOTE method was used to handle imbalanced datasets. Additionally, the binary cross-entropy loss function was employed due to its effectiveness in distinguishing between two classes by measuring the difference between the predicted probabilities and actual binary labels (Creswell et al. 2017). Early stopping was implemented with a patience of 15 epochs, meaning the training process would terminate early if no improvement in validation accuracy were observed over 15 consecutive epochs. This strategy, combined with the use of SMOTE, helped avoid overfitting, improved model performance on imbalanced data and reduced the computational cost by preventing unnecessary training cycles.

During training, the model's performance was monitored on the validation set using metrics such as accuracy, precision, recall, Area Under the Curve (AUC), and F1-score. These metrics provided a thorough evaluation of the model's ability to classify both positive and negative classes, ensuring balanced performance across different clinical conditions in the ECG data. The final model weights were saved based on the best validation performance. This model was then evaluated on a separate test set to confirm its robustness and applicability. The training pipeline was modular, allowing for easy adjustments to hyperparameters and architecture for future research in ECG signal classification using deep learning.

The performance metrics are defined as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1 - score = \frac{2\times Precision\times Recall}{Precision+Recall}$$

$$AUC = \int_{0}^{1} TPR(FPR)\, dFPR$$

Accuracy measures the overall proportion of correctly classified instances, calculated as the ratio of True Positives (TP) and True Negatives (TN) to the total instances. Precision is the ratio of True Positives (TP) to all predicted positives (TP + False Positives, FP), reflecting the model’s ability to avoid false positives. Recall, or Sensitivity, is the ratio of True Positives (TP) to all actual positives (TP + False Negatives, FN), indicating the model’s effectiveness in identifying positive instances. The F1-score, the harmonic mean of Precision and Recall, balances these two metrics. AUC (Area Under the Curve) measures the area under the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various thresholds, assessing the model’s ability to differentiate between classes.

5 Results

In this study, several machine learning models and deep learning methods were employed to evaluate their performance in classifying arrhythmia. The models assessed include Random Forest, Logistic Regression, SVM, MLP, and CNN. For a more in-depth evaluation, feature selection techniques were applied to the machine learning models. The feature used for the experiment was Fourier Transform (marked as FT on 9).

Figure 8 compares the performance of various machine learning models across three key metrics: Accuracy, F1 score, and Area Under the Curve (AUC). The CNN model demonstrates the highest performance across all three evaluation metrics, achieving an average accuracy of 89.29%, an F1 score of 85.69%, and an AUC of 87.98%. These results highlight the model’s superior ability to accurately classify the data while maintaining a strong balance between precision and recall, as reflected by the high F1 score. The model’s near-perfect AUC further underscores its effectiveness in distinguishing between classes, making it a robust choice for this classification task.

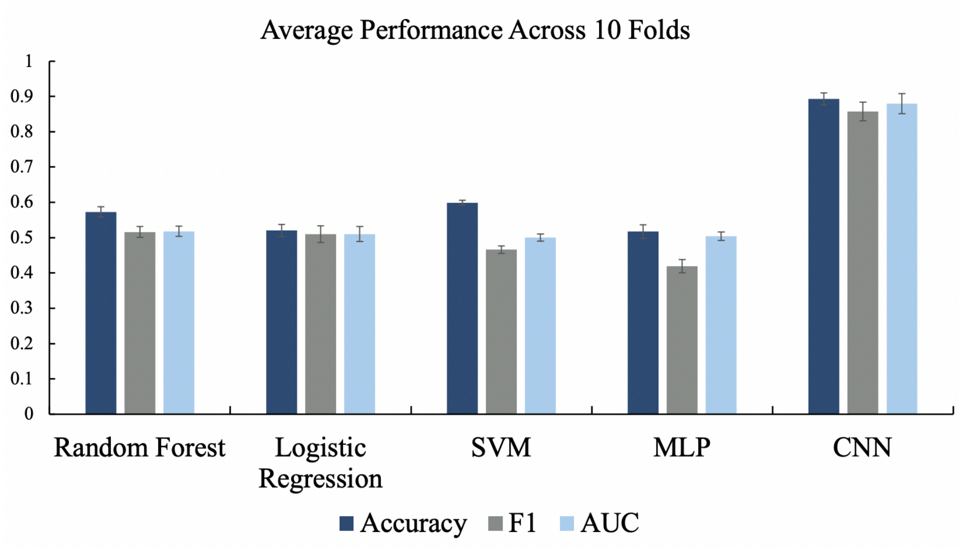


Figure 8. Comparison of Accuracy, F1 Score, and AUC for different machine learning models and feature selection techniques.

Model	**Accuracy (%)**	**F1 Score (%)**	**AUC (%)**
Random Forest	57.3 ± 1.42	51.6 ± 1.56	51.8 ± 1.47
Logistic Regression	52.0 ± 1.76	51.0 ± 2.38	51.0 ± 2.12
SVM	59.9 ± 0.70	46.6 ± 1.02	50.0 ± 1.02
Random Forest (FT)	54.7 ± 1.70	35.8 ± 2.82	51.8 ± 1.55
Logistic Regression (FT)	53.6 ± 1.26	44.2 ± 2.15	52.2 ± 1.40
SVM (FT)	59.9 ± 0.74	57.1 ± 0.32	50.0 ± 0.94
MLP	51.7 ± 1.91	41.9 ± 1.86	50.4 ± 1.18
CNN	89.3 ± 1.66	85.7 ± 2.65	87.9 ± 2.86

Table 1. Performance Metrics for Different Models

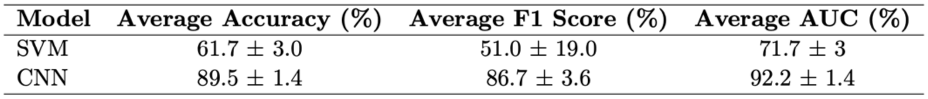
In contrast, both versions of the Random Forest model (with and without feature selection) exhibit moderate performance. The Random Forest model without feature selection achieves an average accuracy of 52%, an F1 score of 51%, and an AUC of 51%, indicating that while it performs marginally better than random guessing, it struggles to balance precision and recall effectively. The version utilizing Fourier Transform (Equation shown by 6) feature selection shows slight improvements in accuracy, with an average of 54.7%, and a modest increase in AUC to 51.8%. However, the F1 score drops significantly to 35.8%, suggesting that this model struggles to consistently maintain a good balance between precision and recall.

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi kn/N}$$

The SVM model exhibits similar performance with and without the application of Fourier Transform (FT) feature selection. Both versions achieve an identical average accuracy of 59.9% and an AUC of 50%, indicating that neither model excels at distinguishing between classes. However, the F1 score, which measures the balance between precision and recall, shows a noticeable improvement when the Fourier Transform is applied. The F1 score increases from 46.6% without FT to 57.1% with FT, suggesting that feature selection using Fourier Transform helps improve the model's ability to handle imbalanced data by better balancing precision and recall. Despite this improvement, the lack of change in AUC indicates that the model's overall discriminatory power remains limited.

The performance of the Logistic Regression model shows mixed results when comparing the version with and without Fourier Transform (FT) feature selection. Without FT, the model achieves an accuracy of 52%, an F1 score of 51%, and an AUC of 51%, indicating a relatively balanced performance across these metrics. When FT is applied, there is a slight increase in accuracy to 53.6% and a small improvement in AUC to 52.2%, suggesting marginally better overall classification ability. However, the F1 score drops to 44.2%, indicating a decline in the model's ability to balance precision and recall. This suggests that while FT improves the model's capacity to differentiate between classes (as seen in the AUC), it does so at the expense of balancing the trade-off between precision and recall.

The MLP model demonstrates performance similar to most machine learning models, with an average accuracy of 51.7%, an F1 score of 41.9%, and an AUC of 50.4%. These values indicate that the MLP model is only slightly better than random guessing in terms of accuracy and AUC, reflecting limited discriminatory power between classes. The relatively low F1 score further suggests that the model struggles to maintain a good balance between precision and recall, possibly due to overfitting or a lack of adaptability to the specific features of the dataset. Although MLP is typically effective for non-linear problems, the results here indicate that it may not be the best model for this classification task without further optimization.



Model	Average Accuracy (%)	Average F1 Score (%)	Average AUC (%)
SVM	61.7 ± 3.0	51.0 ± 19.0	71.7 ± 3
CNN	89.5 ± 1.4	86.7 ± 3.6	92.2 ± 1.4

Table 2. Interpersonal Performance Metrics with Standard Deviation

To test both models, the original dataset was modified to include only segments from patient 100. The new dataset contained 1,666 total segments with 585 abnormal segments. As presented in Table 2 The interpersonal performance metrics reveal significant differences between the SVM and CNN models. The SVM achieved an average accuracy of 61.7%, an average F1 score of 51.0%, and an average AUC of 71.7%. In contrast, the CNN model outperformed the SVM with an impressive average accuracy of 89.5%, an average F1 score of 86.7%, and an AUC of 92.2%.

6 Discussion

6.1 Analysis of Results

The results highlight the distinct advantages of using CNNs for ECG classification com- pared to traditional machine learning models like SVM, Random Forest, and Logistic Regression. CNNs excel in automatically learning and extracting relevant features from raw ECG data, enabling them to capture complex patterns and variations crucial for accurately identifying different types of arrhythmias. CNNs leverage convolutional layers to automatically extract spatial and temporal features from ECG waveforms, eliminating the need for manual feature engineering, which is often required in traditional machine learning approaches. This capability allows CNNs to learn hierarchical representations of ECG signals, where early layers detect basic waveform components such as P-waves and QRS complexes, while deeper layers identify more complex arrhythmic patterns. Unlike SVMs and Random Forest classifiers, which rely on predefined features and may struggle with variations in ECG morphology, CNNs can adapt to diverse waveform structures, making them particularly effective in detecting subtle and rare cardiac anomalies (Salehi et al., 2023).

6.2 Limitations

Although these results demonstrate the potential of CNNs in arrhythmia classification, several challenges remain that must be addressed before deep learning models can be widely adopted in clinical settings. These challenges primarily stem from dataset constraints and model limitations, which impact generalizability and real-world applicability.

One major limitation is the need for large, high-quality datasets to effectively train deep learning models. Neural networks have many parameters, and it is recommended that they have at least ten times more samples than the number of parameters to generalize well. However, acquiring sufficiently large and diverse datasets in the healthcare domain remains challenging. Collecting and annotating medical data is resource-intensive, requiring expert labeling that is often inconsistent. ECG data is further affected by motion artifacts, poor sensor placement, and environmental variations, which compromise data quality. These inconsistencies in labeling and signal integrity introduce additional challenges for training robust deep learning models.

The dataset used in this study, the MIT-BIH Arrhythmia Database, has certain limitations. The database contains a highly imbalanced distribution of heartbeat classes, with the majority being normal heartbeats. This imbalance can lead to biased models that perform well on normal beats but struggle to detect abnormal arrhythmias. Although techniques such as SMOTE were applied to mitigate this issue, they do not introduce new physiological variations that occur in real-world ECG signals.

Another limitation is the dataset's lack of diversity. The MIT-BIH Arrhythmia Database includes recordings from only 48 patients, which restricts its ability to generalize to broader populations. The variations in ECG signals due to age, ethnicity, and health conditions may not be fully captured by this dataset. Future work should focus on incorporating ECG data from multiple sources and diverse demographics to improve model robustness and real-world applicability.

6.3 Future Research

Future research could focus on utilizing larger and more diverse datasets that include ECG recordings from different demographics, clinical conditions, and noise levels. Incorporating data from multiple sources would help improve the model's robustness across different clinical environments. Furthermore, while this study used a CNN-based architecture, future studies could investigate the use of RNNs to generalize longer segments of ECG recordings and use CNN to extract features from smaller segments.

Additionally, transfer learning could be explored as a method to enhance model generalization and performance, particularly in scenarios where labeled ECG data is limited (Gu et al. 2023). By leveraging pre-trained deep learning models trained on large biomedical datasets, researchers could fine-tune CNN architectures for ECG classification, reducing training time while maintaining high accuracy (Salehi et al., 2023, p.5930).

Another promising direction is the integration of multi-modal deep learning approaches by combining ECG with other physiological signals, such as photoplethysmography (PPG) and arterial blood pressure (ABP) [29]. Multi-modal models have been shown to im- prove classification accuracy, reduce false alarms, and enhance robustness to signal artifacts, making them particularly valuable in clinical settings. By leveraging multiple data sources, future studies could improve model reliability and provide a more comprehensive understanding of cardiac activity (Kalidas, 2016, p.1253).

Another critical direction for future research is improving the interpretability of deep learning models through Explainable AI (XAI) techniques. While CNNs have demonstrated high accuracy in ECG classification, their black-box nature remains a significant barrier to clinical adoption. Clinicians require transparency in model decision-making to trust AI-driven diagnoses, particularly in high-risk applications. Future studies could explore post hoc methods like Grad-CAM and SHAP to highlight key ECG features in model predictions. Additionally, integrating attention mechanisms or self-explainable architectures could enhance the interpretability of deep learning decisions. Enhancing interpretability is essential for increasing clinician trust, improving model validation, and facilitating regulatory approval for AI-assisted medical diagnostics (Chaddad et al., 2023, p.634).

7 Conclusion

This research provides a comprehensive approach to using deep learning techniques for arrhythmia classification, demonstrating the significant potential of these methods in medical diagnostics. Among the models tested, the CNN model stood out, achieving the highest performance across all evaluation metrics, with an average accuracy of 89.29%, an F1 score of 85.69%, and an AUC of 87.98%. These results highlight the model's superior ability to accurately classify the data while maintaining a strong balance between precision and re- call. The study's methodologies and findings contribute to the evidence supporting the integration of deep learning models into clinical practice, which could improve patient outcomes through earlier detection of cardiac abnormalities. By automating the extraction of features from ECG data, these models can enhance the accuracy and speed of arrhythmia detection, ultimately leading to better patient care. The modular nature of the training process also allows for adaptability in various clinical settings, making the framework developed in this study applicable to other medical specialties.

In addition to improving classification performance, deep learning also holds promise for the discovery of novel biomarkers, as highlighted in recent research in other fields, such as breast cancer histopathology. By leveraging deep learning algorithms to analyze complex and high-dimensional data, it is possible to uncover biological markers that may not be easily identifiable through traditional methods. In breast cancer, for example, deep learning has been successfully applied to identify new prognostic markers by extracting patterns from histopathology images, as well as linking genomic and proteomic data with clinical outcomes. This capability to discover new biomarkers can significantly enhance personalized medicine by identifying specific indicators of disease progression and treatment response (Mandair et al., 2023, p.21). This promotes the broader adoption of machine learning in healthcare, further advancing diagnostic accuracy and patient outcomes, reducing diagnostic errors, and streamlining treatment decisions. Ultimately, the application of these models could lead to earlier disease detection, more precise interventions, and a significant reduction in healthcare costs by improving efficiency and effectiveness in medical diagnostics.

References

World Health Organization, “The top 10 causes of death,” https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, 2020, [Online; accessed 28- August-2024].

Y. P. Sai et al., “A review on arrhythmia classification using ecg signals,” in 2020 IEEE International Students’ Conference on Electrical, Electronics and Computer Science (SCEECS). IEEE, 2020, pp. 1–6.

M. G. Tsipouras, D. I. Fotiadis, and D. Sideris, “An arrhythmia classification system based on the rr-interval signal,” Artificial intelligence in medicine, vol. 33, no. 3, pp. 237–250, 2005.

E. J. d. S. Luz, T. M. Nunes, V. H. C. De Albuquerque, J. P. Papa, and D. Menotti, “Ecg arrhythmia classification based on optimum-path forest,” Expert Systems with Applications, vol. 40, no. 9, pp. 3561–3573, 2013.

P. Shimpi, S. Shah, M. Shroff, and A. Godbole, “A machine learning approach for the classification of cardiac arrhythmia,” in 2017 International Conference on Computing Methodologies and Communication (ICCMC). IEEE, July 2017, pp. 603–607.

S. Levy, “Diagnostic approach to cardiac arrhythmias,” Journal of cardiovascular pharmacology, vol. 17, p. S32, 1991.

C. Ward, “Reading the electrocardiogram,” Journal of Paediatrics and Child Health, vol. 51, no. 5, pp. 473–475, 2015.

R. Tadros, R. Coronel, and C. R. Bezzina, “Dissecting the genetic basis of the ecg as a means of understanding mechanisms of arrhythmia,” p. e001858, 2017.

N. G. Maity and S. Das, “Machine learning for improved diagnosis and prognosis in healthcare,” in 2017 IEEE aerospace conference. IEEE, 2017, pp. 1–9.

C. U. Kumari, A. S. D. Murthy, B. L. Prasanna, M. P. P. Reddy, and A. K. Pan- igrahy, “An automated detection of heart arrhythmias using machine learning tech- nique: Svm,” Materials Today: Proceedings, vol. 45, pp. 1393–1398, 2021.

L. Jiao and J. Zhao, “A survey on the new generation of deep learning in image processing,” Ieee Access, vol. 7, pp. 172 231–172 263, 2019.

A. Ouelli, B. Elhadadi, and B. Bouikhalene, “Multivariate autoregressive modeling for cardiac arrhythmia classification using multilayer perceptron neural networks,” in 2014 International Conference on Multimedia Computing and Systems (ICMCS). IEEE, 2014, pp. 402–406.22

A. Rajkumar, M. Ganesan, and R. Lavanya, “Arrhythmia classification on ecg using deep learning,” in 2019 5th international conference on advanced computing & communication systems (ICACCS). IEEE, 2019, pp. 365–369.

S. Sahoo, M. Dash, S. Behera, and S. Sabut, “Machine learning approach to detect cardiac arrhythmias in ecg signals: A survey,” Irbm, vol. 41, no. 4, pp. 185–194, 2020.

Z. F. M. Apandi, R. Ikeura, and S. Hayakawa, “Arrhythmia detection using mit-bih dataset: A review,” in 2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA). IEEE, 2018, pp. 1–5.

M. Sampson and A. McGrath, “Understanding the ecg part 2: Ecg basics,” British Journal of Cardiac Nursing, vol. 10, no. 12, pp. 588–594, 2015.

G. S. Mihov and D. H. Badarov, “Application of a reduced band-pass filter in the extraction of power-line interference from ecg signals,” in 2020 XXIX International Scientific Conference Electronics (ET). IEEE, 2020, pp. 1–4.

O. Kwon, J. Jeong, H. B. Kim, I. H. Kwon, S. Y. Park, J. E. Kim, and Y. Choi, “Electrocardiogram sampling frequency range acceptable for heart rate variability analysis,” Healthcare informatics research, vol. 24, no. 3, pp. 198–206, 2018.

R. Blagus and L. Lusa, “Smote for high-dimensional class-imbalanced data,” BMC bioinformatics, vol. 14, pp. 1–16, 2013.

J. A. Nasiri, M. Naghibzadeh, H. S. Yazdi, and B. Naghibzadeh, “Ecg arrhythmia classification with support vector machines and genetic algorithm,” in 2009 Third UKSim European symposium on computer modeling and simulation. IEEE, 2009, pp. 187–192.

A. Mammone, M. Turchi, and N. Cristianini, “Support vector machines,” Wiley In- terdisciplinary Reviews: Computational Statistics, vol. 1, no. 3, pp. 283–289, 2009.

L. Breiman, “Random forests,” Machine learning, vol. 45, pp. 5–32, 2001.

H. Ramchoun, Y. Ghanou, M. Ettaouil, and M. A. Janati Idrissi, “Multilayer perceptron: Architecture optimization and training,” 2016.

M. Krichen, “Convolutional neural networks: A survey,” Computers, vol. 12, no. 8, p. 151, 2023.

N. Bjorck, C. P. Gomes, B. Selman, and K. Q. Weinberger, “Understanding batch normalization,” Advances in neural information processing systems, vol. 31, 2018.

N. Murray and F. Perronnin, “Generalized max pooling,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 2473–2480.23

A. Creswell, K. Arulkumaran, and A. A. Bharath, “On denoising autoencoders trained to minimise binary cross-entropy,” arXiv preprint arXiv:1708.08487, 2017.

A. W. Salehi, S. Khan, G. Gupta, B. I. Alabduallah, A. Almjally, H. Alsolai, T. Sid- diqui, and A. Mellit, “A study of cnn and transfer learning in medical imaging: Ad- vantages, challenges, future scope," Sustainability, vol. 15, no. 7, p. 5930, 2023.

X. Gu, F. Deligianni, J. Han, X. Liu, W. Chen, G.-Z. Yang, and B. Lo, "Beyond supervised learning for pervasive healthcare," IEEE Reviews in Biomedical Engineering, vol. 17, pp. 42–62, 2023.

V. Kalidas and L. Tamil, "Cardiac arrhythmia classification using multi-modal signal analysis," Physiological measurement, vol. 37, no. 8, p. 1253, 2016.

A. Chaddad, J. Peng, J. Xu, and A. Bouridane, "Survey of explainable ai techniques in healthcare," Sensors, vol. 23, no. 2, p. 634, 2023.

D. Mandair, J. S. Reis-Filho, and A. Ashworth, "Biological insights and novel biomarker discovery through deep learning approaches in breast cancer histopathology," NPJ breast cancer, vol. 9, no. 1, p. 21, 2023.