Research Article

Received: 3 January 2024 / Accepted: 7 April 2024 / Published: 5 May 2024

Classification of ECG Signals Using Machine Learning Techniques

Diego Fernando Sendoya Losada

Julián José Soto Gómez

Julián Andrés Zúñiga Vela

*Electronics Engineering Program,
Faculty of Engineering,
Universidad Surcolombiana,
Colombia*

DOI: https://doi.org/10.36941/ajis-2024-0067

Abstract

Cardiovascular diseases are one of the leading causes of mortality in contemporary society. With the growth in the accumulation of medical data, new opportunities have arisen to enhance diagnostic accuracy using machine learning techniques. Heart diseases present symptoms that can be similar to other disorders or be mistaken for signs of aging. Furthermore, diagnosing based on electrocardiogram (ECG) signals can be challenging due to the variability in signal length and characteristics. This article has developed a methodology for classifying ECG signals using the k-Nearest Neighbor (kNN) algorithm and statistical techniques. 9000 ECG signal samples from the PhysioNet database were processed. The signals were normalized to a length of 9000 samples, and relevant features for classification, such as median, standard deviation, skewness, among others, were extracted. Multiple kNN models with different parameters were trained and evaluated on a test set. The models exhibited high performance in classifying normal signals but faced difficulties in correctly classifying signals with arrhythmias. The weighted kNN algorithm demonstrated the best accuracy, although all models showed a tendency to misclassify abnormal signals due to data imbalance. While significant accuracy was achieved in ECG signal classification, there is still room for improvement. Future strategies could involve extracting more relevant features, addressing data imbalance, and fine-tuning model hyperparameters. Integrating domain knowledge from the medical field and advanced signal processing techniques could further enhance classification accuracy.

Keywords: *Arrhythmias, Cardiovascular diseases, ECG, kNN, Machine Learning*

1. Introduction

Cardiovascular diseases are one of the main causes of mortality in contemporary society. With the increase in the accumulation of medical data, health professionals find new possibilities to improve diagnostic accuracy. Recently, the use of informatics in clinical decision-making has been refined.

Heart conditions comprise a variety of disorders that affect the heart and its components, including arteries, muscles, valves, and internal electrical circuits responsible for muscle contraction. According to the Centers for Disease Control and Prevention, heart disease remains the leading cause of death in countries including India, the United Kingdom, the United States, Canada, and Australia. Globally, cardiovascular diseases (CVD) represent 31% (equivalent to 17.9 million) of all annual deaths, imposing a considerable clinical (mortality and disability), health, and economic burden. In the United States, for example, heart disease is responsible for one in four deaths (Thomas & Princy, 2016).

In most nations, heart disease affects men and women equally. This underscores the importance of considering the risk factors associated with cardiovascular disease. Although there is a genetic component, certain lifestyle variables significantly influence the development of heart conditions. Known risk factors include malnutrition, hypertension, high blood cholesterol levels, diabetes, obesity, a sedentary lifestyle, stress, and poor hygiene (Saboor et al., 2022). These are some of the factors that increase the probability that a patient will develop CVD.

However, the diagnosis of these diseases becomes complicated for health professionals since the symptoms are often similar to those of other conditions or can be confused with signs of aging. In this context, the New Technologies Research Group of Universidad Surcolombiana has focused its research efforts on exploring various machine learning techniques. The main purpose is to achieve reliable and accurate detection of cardiac abnormalities, a highly relevant line of research given the prevalence and seriousness of cardiovascular conditions.

Machine learning faces significant challenges when handling large-dimensional data sets (Dogan et al., 2021). An effective strategy to address these difficulties is feature weighting, which allows the reduction of redundant data and decreases processing time, thus improving the performance of the algorithms. This approach is precious since analyzing a large set of features may require a large amount of memory and result in overfitting.

Areas such as healthcare management, genomic expression, medical imaging, and the Internet of Things (IoT) are often characterized by small but highly informative feature sets. In this way, dimensionality reduction plays a crucial role: feature extraction allows data to be transformed and simplified, while feature selection focuses on decreasing the complexity of the data set by removing redundant features (Spencer et al., 2020).

In the study by Aggrawal and Pal (2020), a sequential attribute selection strategy was implemented with the aim of identifying the most critical factors and detecting instances of mortality in patients with cardiovascular disease during treatment. Several machine learning approaches, including linear discriminant analysis, k-nearest neighbors' algorithm, random forest, and decision trees, were employed in their methodology. The effectiveness of each model was measured using various performance metrics. According to the results presented, the random forest approach achieved an accuracy of 86.67%, indicating that it was the most effective of the techniques evaluated in their study.

Gao et al. (2021) proposed a model for the prediction of cardiac diseases. They used an approach based on ensemble methods and feature extraction techniques, applied to the Cleveland heart disease data set. The study results indicate that the implementation of the bagging sets strategy, in combination with Decision Trees (DT) and Principal Component Analysis (PCA), offered superior performance compared to other evaluated methods.

Takci (2018) implemented feature selection methods and classification algorithms of various categories to predict heart attacks. Different parameters were evaluated, including model accuracy, processing speed, and Receiver Operating Characteristic (ROC) curve analysis results. The results showed that the maximum accuracy obtained was 82.59% without feature selection and 84.81% with feature selection. To achieve a model accuracy of 84.81%, Naive Bayes and linear Support Vector Machines (SVM) algorithms were used.

Latha and Jeeva (2019) proposed a combined technique of feature selection and ensemble classification for heart disease risk prediction. The results of the study evidenced that ensemble-based approaches, such as Bagging and Boosting, can accurately predict heart disease risk, in addition to improving the prediction accuracy of weak classifiers.

Garate-Escamila et al. (2020) applied Principal Component Analysis (PCA) and Chi-square test to develop a hybrid dimensionality reduction approach. For their research, they selected data sets from Hungary, Cleveland, and a combination of both. In a subsequent classification process, Random Forests, Gradient Boosted Trees, Decision Trees, Multilayer Perceptron, and Logistic Regression were employed. The hybrid approach based on PCA and Chi-square test, implemented with Random Forests, proved to be the most accurate, achieving 98.7% accuracy for the Cleveland dataset, 99.0% for the Hungary dataset, and 99.4% for the combined Cleveland-Hungary dataset.

Spencer et al. (2020) proposed a model using dimensionality reduction techniques, including Principal Component Analysis (PCA), Chi-square, Relief, and Symmetric Uncertainty tests. The authors also discussed in their paper the advantages of feature selection when applied to cardiac datasets, highlighting the effectiveness of these techniques in the environment of cardiac disease detection and prediction.

Senan et al. (2021) developed a method for the diagnosis of chronic kidney disease, in which they employed Recursive Feature Elimination to determine the most relevant features in this clinical context. In the study, several classification techniques were used: K-nearest neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest. The efficacy of each of these models was evaluated and it was found that the Random Forest offered the highest accuracy, reaching 100% efficacy in predicting CKD.

Almansour et al. (2019) proposed a model capable of predicting chronic kidney disease. For their research, they resorted to the dataset consisting of 400 instances, coming from the UCI repository. They employed Support Vector Machines and Artificial Neural Networks as classification techniques. The effectiveness of the classifiers was evaluated through the selection of optimal features and measurement of model training time. The effectiveness of each classification technique was compared considering both model accuracy and computational efficiency.

The purpose of this research is to develop a methodology for the classification of human electrocardiogram (ECG) signals using Machine Learning (ML) techniques. The procedure to follow comprises the extraction of clinically relevant features and the subsequent training of several ML models, with the purpose of recognizing the following cardiac rhythms:

- Normal Sinus Rhythm (N)
- Arrhythmia (A)

The ECG data used in this study were extracted from the PhysioNet Challenge 2017 database, available for consultation at the following URL: https://physionet.org/challenge/2017/.

2. Materials and Methods

2.1 Materials

The k-Nearest Neighbor (kNN) algorithm is a supervised classification method within the field of machine learning. This algorithm classifies objects based on the closest data points in the training data set. The 'k' in kNN is a parameter that refers to the number of nearest neighbors to be considered during the classification of a new object.

The kNN classification operates on the principle that objects in a multidimensional space tend to be closest to objects of the same class. When a new query point is presented, this algorithm measures its distance to all other points in the training data set. Distances can be calculated using various metrics, the most common being Euclidean distance, Manhattan, and Minkowski, among others. Subsequently, the algorithm identifies the 'k' nearest points, i.e., their 'k' nearest neighbors.

Finally, the kNN algorithm assigns to the new query point the class that is most frequent among its 'k' nearest neighbors. If k=1, then the object is simply assigned to the class of that single nearest neighbor. In the case where there is a tie, i.e., two or more classes have the same number of representatives among the 'k' nearest neighbors, the nearest neighbor (among the 'k') can be taken as the tiebreaker.

Because of its simplicity and efficiency, the kNN algorithm is frequently used as a benchmark in machine learning research and remains a popular method in numerous applications, from pattern recognition and data mining to machine learning and computer vision.

Classifiers based on the k-Nearest Neighbor (kNN) algorithm depend on the choice of 'k', i.e., the number of neighbors considered for the classification of a new object. The value of 'k' directly influences the classification accuracy, and its optimal choice may vary depending on the specific problem and data distribution.

Fine KNN: This classifier considers only the nearest neighbor (k=1) for classification. It tends to make very detailed distinctions between classes but can be susceptible to noise and data anomalies, resulting in high variance and possibly overfitting.

Mean KNN: This classifier considers the 10 nearest neighbors (k=10) for classification. It makes more general distinctions between classes, providing a balance between bias and variance.

Coarse KNN: This classifier considers the 100 nearest neighbors (k=100) for classification. It makes even more general distinctions between classes. This approach can be useful when classes are widely overlapping, although it can introduce high bias.

KNN Cosine: This classifier uses cosine distance to measure the similarity between objects. It is particularly useful when working with data in text format or when the direction of the data is more important than its absolute magnitude.

Cubic KNN: This classifier uses a cubic distance metric. It is less common than the Euclidean or cosine metrics but can be useful in certain problems with a particular data structure.

Weighted KNN: This classifier weights the contribution of each neighbor based on its distance to the query object. Closer neighbors will have a greater influence on the classification than those farther away. This approach can be useful when class similarity is expected to decrease with distance.

2.2 Data Preprocessing

From the database, 5788 ECG signals with their respective labels are extracted and quantified to determine the number of signals with Normal Sinus Rhythm (N) and with Arrhythmia (A). It is observed that there are 5050 ECG signals labeled with N and 738 labeled with A. Preliminarily, it is identified that the length of the signals is not constant, so a histogram reflecting the length of the signals is generated.

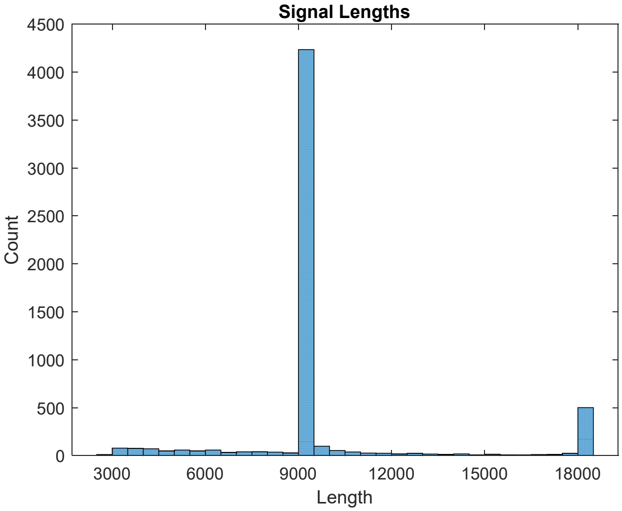


Figure 1: Electrocardiogram signal length

It is important to note that, although most ECG signals have a length of 9000 samples, there is appreciable variability. Since ECG signals exhibit varying lengths, it is necessary to normalize them so that they all possess a length of 9000 samples. Therefore, signals containing less than 9000 samples are excluded. In case a signal has more than 9000 samples, it is divided into as many 9000-sample segments as possible, discarding the excess samples. For example, a signal with 18500 samples will be split into two 9000-sample signals, ignoring the remaining 500 samples.

This results in a new database with 5655 ECG signals with a length of 9000 samples, as can be verified with another histogram. It is now observed that there are 4937 N-labeled ECG signals and 718 A-labeled ECG signals.

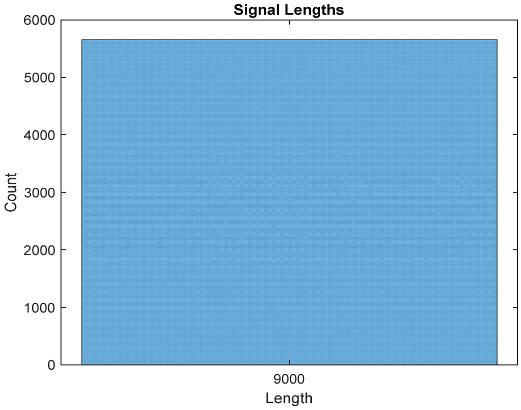


Figure 2: Normalization of the length of a signal

A segment of a signal corresponding to each class can be displayed.

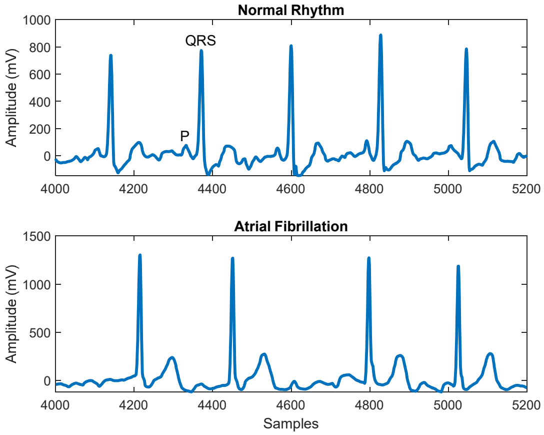


Figure 3: Segments of a normal signal and with atrial fibrillation

Two common characteristics of signals classified as Arrhythmia (A) are:

- Their pulses present irregular intervals.
- They lack a P wave, which precedes the QRS complex in normal heartbeats.

2.3 Feature Extraction

Machine Learning (ML) models use features (also known as predictors) as input, which in terms of statistical modeling are called dependent variables, and generate a predicted value as output, known as the independent variable. The features can be external measurements or any other value that is related to the output variable. The latter can be thought of as internal characteristics that are generated through the transformation of the original data.

The extraction of the most appropriate and valuable features requires a thorough knowledge of the application domain. The use of statistical measures of the data can be an excellent starting point for feature extraction. The best features are considered those features that show a considerable difference between the different classes.

Initially, the signals corresponding to Normal Sinus Rhythm (N) are separated from the signals corresponding to Arrhythmia (A), and the distribution of signals from both categories is observed by plotting a histogram of one signal from each class. The median (MDN), standard deviation (SD), skewness (SKEW), and kurtosis (K) of both signals are also calculated and displayed.

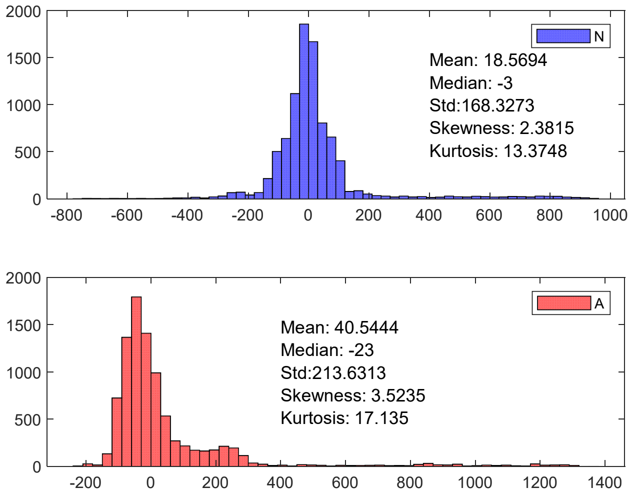
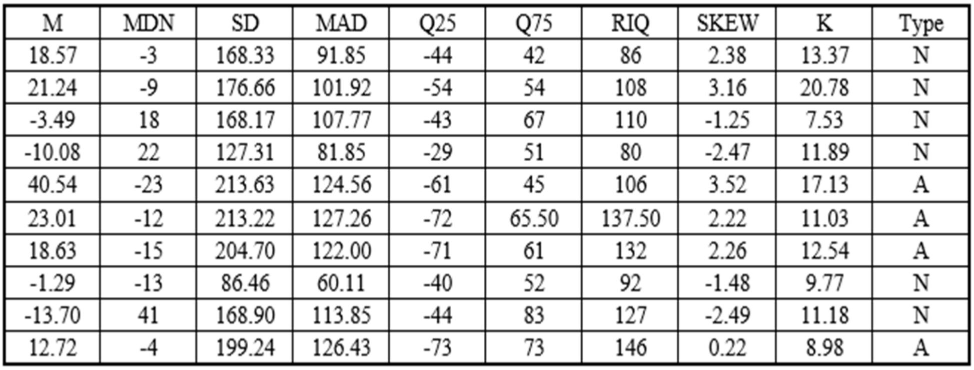


Figure 4: Distributions and characteristics of a type N and type A signals

From the analysis of the statistical variables, it can be observed that the signals (N and A) present significantly different characteristics. For this reason, these characteristics will be selected. Additionally, the mean absolute deviation (MAD), the 25th (Q25) and 75th (Q75) quantiles, and the interquartile range (IQR) of the signal will be included. Finally, the features are calculated for the entire data set and a table containing the 10 statistical measures of the signals is generated. The first 10 values of this table are presented below:

Table 1: Results of statistical variables



M	MDN	SD	MAD	Q25	Q75	RIQ	SKEW	K	Type
18.57	-3	168.33	91.85	-44	42	86	2.38	13.37	N
21.24	-9	176.66	101.92	-54	54	108	3.16	20.78	N
-3.49	18	168.17	107.77	-43	67	110	-1.25	7.53	N
-10.08	22	127.31	81.85	-29	51	80	-2.47	11.89	N
40.54	-23	213.63	124.56	-61	45	106	3.52	17.13	A
23.01	-12	213.22	127.26	-72	65.50	137.50	2.22	11.03	A
18.63	-15	204.70	122.00	-71	61	132	2.26	12.54	A
-1.29	-13	86.46	60.11	-40	52	92	-1.48	9.77	N
-13.70	41	168.90	113.85	-44	83	127	-2.49	11.18	N
12.72	-4	199.24	126.43	-73	73	146	0.22	8.98	A

2.4 Data Preparation for Training

To enable an efficient evaluation of the predictive models, the data set is divided into two distinct subsets: a training set and a test set. The training set is used to train the models, while the test set is used after the training phase to examine the performance of the models when confronted with data not previously exposed to the training process.

It is recommended that 60% to 90% of the total data be assigned to training, leaving the remainder for the testing procedure. In this work, 80% of the data is used for training and the remaining 20% for testing. In the training set, the number of signals corresponding to Normal Sinus Rhythm (N) is 3942 and the number of signals corresponding to Arrhythmia (A) is 582. On the other hand, the test set contains 995 N signals and 136 A signals.

3. Results

In the field of Machine Learning (ML), one of the most significant challenges is the abundance of machine learning algorithms that could address the same problem. It is not feasible to predict a priori which of them will provide the most optimal performance.

3.1 Training Data

In order to have a reference of how the training data are distributed, the following graph shows how they are grouped taking the mean value (M) and the median (MDN) as a reference:

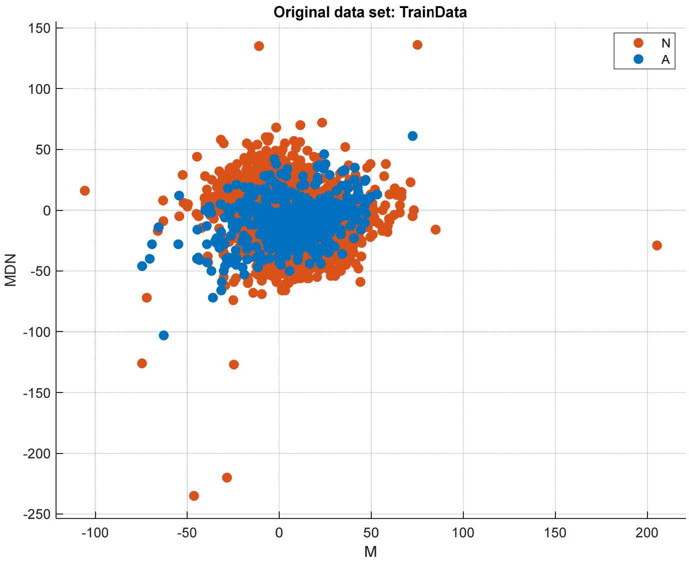


Figure 5: Distribution of trained data

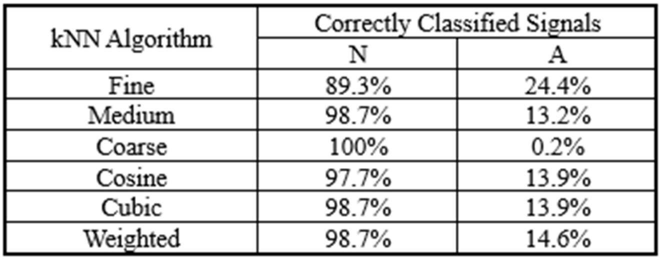
For all kNN algorithms, standardized data were used for training. A uniform distance weight was applied to all algorithms, except for the weighted kNN, which utilized the inverse square. The other parameters used to fine-tune each kNN algorithm, along with the results obtained during training, are summarized in the following table:

Table 2: Data obtained during training

Parameters	kNN Algorithm					
	Fine	Medium	Coarse	Cosine	Cubic	Weighted
Number of Neighbors	1	10	100	10	10	10
Metric Distance	Euclidean	Euclidean	Euclidean	Cosine	Minkowski (cubic)	Euclidean
Accuracy	80.9%	87.7%	87.2%	86.9%	87.8%	87.9%
Total Cost	863	558	581	591	552	549
Prediction Speed	~11000 obs/sec	~8300 obs/sec	~4700 obs/sec	~5500 obs/sec	~2100 obs/sec	~7200 obs/sec
Training Time	13.265 seconds	21.298 seconds	20.525 seconds	19.152 seconds	22.196 seconds	17.382 seconds
Model Size	~503 kB	~503 kB	~503 kB	~398 kB	~503 kB	~503 kB

The highest accuracy during training is achieved with the weighted kNN algorithm. Likewise, when examining the confusion matrix for each algorithm, the following can be observed:

Table 3: Accuracy of algorithms with respect to signals



kNN Algorithm	Correctly Classified Signals	
	N	A
Fine	89.3%	24.4%
Medium	98.7%	13.2%
Coarse	100%	0.2%
Cosine	97.7%	13.9%
Cubic	98.7%	13.9%
Weighted	98.7%	14.6%

The confusion matrix for the weighted kNN algorithm, obtained during training, is presented below:

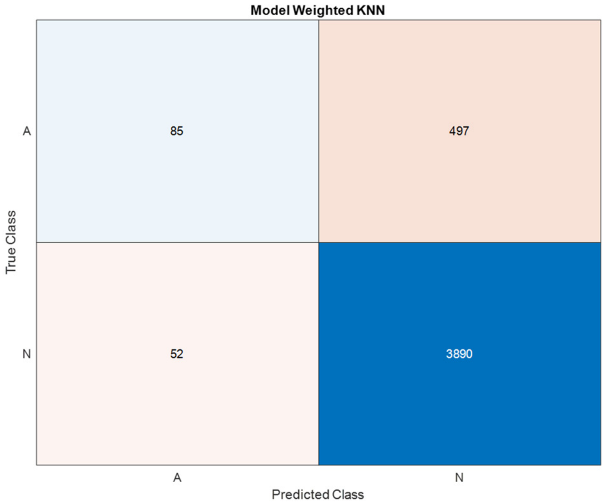


Figure 6: Confusion matrix

Most of the trained models exhibit a high false negative rate in the detection of signals corresponding to Arrhythmia (A). This situation derives from the existing imbalance in the classes, where it can be recalled that the training set includes 582 signals with Arrhythmia (A) versus 3942 signals corresponding to Normal Sinus Rhythm (N). This imbalance causes the model to tend to bias towards the predominant class, in our case, the normal signal.

There are multiple strategies to address the class imbalance. Some examples include over-sampling or under-sampling techniques. An alternative to reduce misclassification is the use of cost-sensitive models, which impose a penalty on the majority class. The search for optimal coefficients for misclassification costs is commonly performed manually through a trial-and-error process.

3.2 Test Data

It is possible to evaluate the performance of each model with the test data set, i.e., data that has not been used during the training phase. This process allows measuring how the models will handle new and unseen data and is a common practice in machine learning to ensure that the models are not overfitted to the training data.

The results obtained with each kNN algorithm during the testing phase are summarized below:

Table 4: Data obtained from test data

kNN Algorithm	Precision	Total Cost	Correctly Classified Signals	
			N	A
Fine	81.1%	214	89.5%	19.1%
Medium	88.3%	132	98.7%	12.5%
Coarse	88.0%	136	100%	0%
Cosine	87.7%	139	97.7%	14.7%
Cubic	88.1%	135	98.4%	12.5%
Weighted	88.2%	134	98.7%	11%

Once again, it is observed that the weighted kNN algorithm exhibits high precision. The confusion matrix for the weighted kNN algorithm, obtained during the testing phase, is presented below:

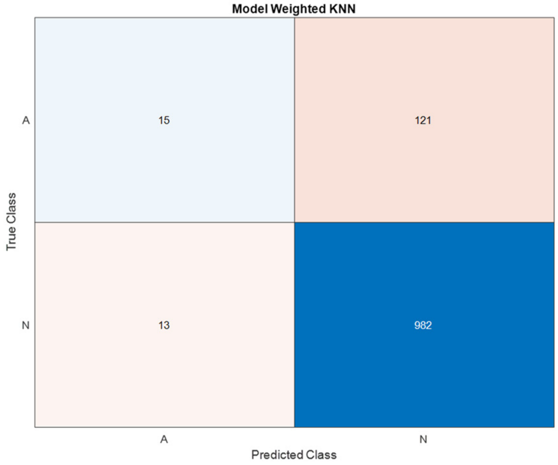


Figure 7: Confusion matrix of test data

One of the most significant challenges identified in this study is the classification of abnormal ECG signals, which is crucial for the accurate diagnosis of arrhythmias. Abnormal signals often exhibit subtle and complex variations, making their identification and classification a non-trivial task for machine learning models. The difficulty is compounded by the inherent data imbalance and the potential for overfitting to the more prevalent normal signal class. To address these challenges, several strategies could be employed. Enhancing the dataset through techniques such as synthetic data generation or oversampling of the minority class can help mitigate the issue of data imbalance. Additionally, incorporating advanced feature extraction methods that can capture the intricate patterns unique to abnormal signals may improve classification performance. Exploring alternative machine learning algorithms known for their robustness in handling imbalanced data, such as ensemble methods or deep learning approaches, could also offer significant benefits. Moreover, the application of cost-sensitive learning, where higher penalties are assigned to misclassifications of the minority class, may further refine the model's ability to distinguish between normal and abnormal signals accurately.

4. Conclusions

This study offers significant insights into the classification of ECG signals using machine learning techniques, with a particular focus on the k-Nearest Neighbor (kNN) algorithm. Despite its contributions, it faces inherent research limitations that necessitate attention for a balanced interpretation of the results. The first notable limitation is data imbalance, where the disproportionate representation of normal signals compared to arrhythmic ones could bias the model towards the majority class, compromising the reliability of arrhythmia detection. Secondly, feature selection is critical to the models' performance, where chosen features based on the statistical properties of ECG signals might not capture all nuances necessary to distinguish between complex arrhythmias. Additionally, the criteria for model evaluation must be carefully considered, as metrics such as precision, recall, and the F1 score can offer a more comprehensive assessment of model capabilities, especially in correctly identifying arrhythmia cases.

The models demonstrated remarkably high performance in classifying normal signals but showed deficiencies in classifying signals with Arrhythmia. Despite developing machine learning models to identify cardiac arrhythmias from ECG waveforms and evaluating their performance, the problem has not been satisfactorily solved with a maximum accuracy of less than 90%.

Considering the challenges and limitations identified throughout this study, providing specific recommendations for future research directions is imperative to enhance the classification of ECG signals using machine learning techniques. Data imbalance remains a critical concern, and future studies could explore innovative resampling techniques or the application of synthetic data generation methods like SMOTE to ensure a more equitable distribution of classes. Moreover, extracting more relevant features is paramount for improving classification accuracy. Advanced signal processing techniques and the incorporation of machine learning methods capable of automatic feature learning, such as deep learning, could uncover subtle patterns in ECG signals that traditional methods might overlook. Integrating domain knowledge is another promising avenue. Collaboration with medical experts can guide the selection of clinically significant features and the development of composite features that better represent the physiological characteristics of arrhythmias. Furthermore, exploring models that incorporate patient history and demographic information could offer a more holistic approach to diagnosis.

References

Aggrawal, R., & Pal, S. (2020). Sequential feature selection and machine learning algorithm-based patient's death events prediction and diagnosis in heart disease. SN Computer Science, 1(6).

Almansour, N. A., Syed, H. F., Khayat, N. R., et al. (2019). Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. Computers in Biology and Medicine, 109, 101–111.

Dogan, O., Tiwari, S., Jabbar, M. A., & Guggari, S. (2021). A systematic review on AI/ML approaches against COVID-19 outbreak. Complex Intelligence Systems, 7, 2655–2678.

Gao, X. Y., Ali, A. A., Hassan, S. H., & Anwar, E. M. (2021). Improving the accuracy for analyzing heart diseases prediction based on the ensemble method. Complexity, 2021, Article ID 6663455, 10 pages.

Garate-Escamila, A. K., Hassani, A. E., & Andres, E. (2020). Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked, 19.

Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked, 16.

Saboor, A., Usman, M., Ali, S., Samad, A., Abrar, M. F., & Ullah, N. (2022). A method for improving prediction of human heart disease using machine learning algorithms. Mobile Information Systems, 2022.

Senan, E. M., Al-Adhaileh, M. H., Alsaade, F. W., et al. (2021). Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. Journal of Healthcare Engineering, 2021, Article ID 1004767, 10 pages.

Spencer, R., Thabtah, F., Abdelhamid, N., & Thompson, M. (2020). Exploring feature selection and classification methods for predicting heart disease. DIGITAL HEALTH, 6.

Takci, H. (2018). Improvement of heart attack prediction by the feature selection methods. Turkish Journal of Electrical Engineering and Computer Sciences, 26, 1–10.

Thomas, J., & Princy, R. (2016). Human heart disease prediction system using data mining techniques. 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), 1-5.