Vol.15, Issue No 2, 2025

HEALTH RISK ASSESSMENT USING MACHINE LEARNING CLASSIFIERS ON WEARABLE IOT DEVICES

Raparla Sai Manogna,
UG Student,
Department of CSE,
St. Martin's Engineering College,
Secunderabad, Telangana, India.
manognaraparla41@gmail.com

Mr. D. Sai Kiran,
Assistant Professor,
Department of CSE,
St. Martin's Engineering College,
Secunderabad, Telangana, India.

Abstract:

Electrocardiogram (ECG) arrhythmia classification is crucial for diagnosing and managing cardiac conditions, as arrhythmias can lead to serious health complications if left untreated. The application of machine learning-based ECG arrhythmia classification has significant implications for cardiac healthcare and clinical practice. Accurate and timely classification of arrhythmias enables healthcare professionals to diagnose cardiac conditions, determine appropriate treatment strategies, and monitor patient health effectively. Moreover, machine learning models can support remote monitoring and telemedicine initiatives, allowing for early detection of arrhythmias and timely intervention, particularly in underserved or remote areas. Additionally, ECG arrhythmia classification facilitates research efforts aimed at understanding the underlying mechanisms of arrhythmias and developing novel diagnostic and therapeutic approaches. Existing methods for ECG arrhythmia classification often rely on manual interpretation by cardiologists or rule-based algorithms, which may be subjective, time-consuming, and prone to errors. These methods may struggle to accurately classify complex arrhythmias or differentiate between similar arrhythmia patterns. Moreover, traditional approaches may overlook subtle changes in ECG signals or fail to capture the full spectrum of arrhythmia features, leading to suboptimal classification performance. Additionally, the increasing volume and complexity of ECG data pose challenges for traditional classification methods, necessitating more advanced and data-driven approaches. The proposed system utilizes machine learning techniques to automate and enhance ECG arrhythmia classification, addressing the limitations of existing methods. This work employs supervised learning algorithms to train models on ECG data and classify different types of arrhythmias. By preprocessing ECG signals, extracting informative features, and leveraging advanced classification techniques. The proposed models can accurately identify arrhythmia patterns and classify them into relevant categories.

Keywords: Health Risk Assessment, Naïve Bayes Classifier, Feature Extraction, KNN(K Nearest Neighbour), Real Time Health Monitoring, SMOTE(Synthetic Minority Oversampling Technique), Model Evaluation, IOT in Healthcare, ECG(Electrocardiogram).

Recent ad the Internet of Things (IoT) have made health monitoring more efficient and accessible. Devices like smartwatches, fitness trackers, and medical sensors are now widely used to collect real-time health data, such as heart rate, blood pressure, oxygen levels, physical activity, and sleep patterns. These devices allow for continuous monitoring, providing detailed insights into a person's health. This information helps identify potential risks early and supports preventive care measures.

Machine learning (ML) has become a key tool in analyzing the large amounts of data generated by these wearable devices. By finding patterns and detecting unusual changes in the data, ML algorithms can predict health risks, identify early signs of issues, and help healthcare professionals make better decisions. This approach allows healthcare to be more personalized, offering solutions tailored to individual needs. The combination of wearable IoT devices and machine learning has significant potential to improve healthcare by enabling constant monitoring, early detection of problems, and reducing costs.

However, challenges like ensuring data quality, privacy, and security must be addressed. Despite these challenges, the benefits of using machine learning in health monitoring are substantial, especially in improving healthcare outcomes and patient well-being. This research focuses on how machine learning classifiers can be applied to health risk assessment using data from wearable IoT devices and their potential to transform healthcare.

A key area where this technology is making a difference is in cardiac health. Wearable devices can monitor heart health in real time and help identify irregular heart rhythms, known as arrhythmias. If left undetected, arrhythmias can lead to severe health issues like strokes or sudden cardiac death. Traditionally, detecting arrhythmias has relied on manual ECG interpretation by cardiologists or rule-based algorithms. However, these methods can be time-consuming, subjective, and not always accurate.

Machine learning offers a more efficient and accurate solution to this problem. With its ability to process large amounts of data and recognize complex patterns, machine learning can automate ECG arrhythmia classification. This paper introduces "Predictive Pulse," a system that uses machine learning to identify and classify arrhythmias with high accuracy. Predictive Pulse can analyze large ECG datasets in real time, making it useful not only for clinical diagnostics but also for public health monitoring.

The development of ECG technology began with the invention of the electrocardiograph by Willem Einthoven in the early 20th century, which enabled the recording of heart activity. While this innovation revolutionized cardiac care, traditional methods still rely heavily on manual interpretation and basic algorithms. These approaches struggle to keep up with the growing volume and complexity of ECG data, especially with the introduction of wearable devices. This gap highlights the need for scalable and efficient solutions like machine learning.

This research aims to address the limitations of traditional methods and improve the accuracy and efficiency of arrhythmia classification. Manual analysis is prone to errors and inconsistencies, while rule-based algorithms often lack the flexibility to detect subtle variations in ECG signals. Machine learning models, trained on extensive datasets, can overcome these challenges by automating the classification process, enabling early intervention and personalized treatments.

The Predictive Pulse system represents a significant advancement in this field. It improves cardiac diagnostics by providing accurate and real-time classification of arrhythmias. The system also supports remote monitoring and telemedicine, making healthcare more accessible, particularly for people in remote areas. Additionally, it is scalable and can be used in large-scale public health initiatives. By combining wearable IoT technology with machine learning, Predictive Pulse has the potential to revolutionize how arrhythmias are detected, diagnosed, and managed, ultimately improving cardiac healthcare and patient outcomes.

II. RELATED WORK

The heart is a vital organ that functions as the body's engine, continuously pumping blood to deliver oxygen and nutrients while removing waste products. Beating approximately 100,000 times daily, the heart generates electrical activity that can be recorded as an electrocardiogram (ECG) using skin electrodes. The ECG captures the electrical signals of the heart in the form of PQRST waveforms, which reveal vital information about heart rate variability (HRV) and morphology, playing a crucial role in the diagnosis of arrhythmias [1]. Cardiac arrhythmias such as premature atrial contraction (PAC) and premature ventricular contraction (PVC) are significant indicators of cardiac health. PACs originate in the atria or atrioventricular (AV) node and are marked by abnormal P wave morphology and compensatory pauses shorter than two normal RR intervals. PVCs arise from the ventricles and lead to a compensatory pause that inhibits the subsequent sinus beat, further emphasizing the importance of timely detection and classification for maintaining cardiac health [2, 3].

The field of healthcare has witnessed groundbreaking advancements with the advent of biomedical sensors, the Internet of Medical Things (IoMT), and artificial intelligence (AI). These innovations have revolutionized traditional healthcare systems by enabling real-time monitoring, improving diagnostic accuracy, and reducing the need for frequent clinical visits [4, 5]. IoMT, when integrated with microelectronics, 5G technology, and AI, forms the foundation of smart healthcare systems, offering cost-effective and precise solutions [6].

The human body is known as a complex electromechanical system generating several types of biomedical signals, such as an electrocardiogram (ECG), which is a record of the dynamic changes of the human body that need to be monitored by smart healthcare systems. For instance, the EKG sensor measures cardiac electrical potential waveforms. It is used to create standard 3-lead electrocardiogram (EKG) tracings to record the electrical activity in the heart or to collect surface electromyography (sEMG) to study the contractions in the muscles of the arm, leg, or jaw. Simply, an ECG graphs heartbeats and rhythms. The classification of an ECG heartbeat plays a substantial role in smart healthcare systems [7,8], where the presence of multiple cardiovascular problems is generally indicated by an ECG. In the subsequent ECG waveform, diseases cause defects. However, early diagnosis via an ECG allows for the selection of suitable cardiac medication and is thus very important and helpful for reducing heart attacks [9].

The method of detecting and classifying arrhythmia is not an easy task and may be very difficult even for professionals because sometimes it is important to examine multiple pulses of ECG data, obtained, for example, during hours, or even days, by a Holter clock. Furthermore, there is a possibility for errors by humans during the ECG recording study due to fatigue. Building a fully automatic arrhythmia detection or classification system is difficult. The difficulty comes from the large amount of data and the diversities in the ECG signals due to the nonlinearity, complexity, and low amplitude of ECG recordings, as well as the nonclinical conditions, such as noise [10]. Despite all these difficulties, methods for ECG arrhythmia classification have been widely explored [11,12] but choosing the best technique for smart patient monitoring depends on the robustness and performance of these methods. Several convolutional neural network (CNN)-based approaches have been introduced for the task [13,14].

In [15], a subject-adaptable ECG arrhythmia classification model was proposed and trained with unlabeled personal data. It achieved an average performance of 99.4% classification accuracy on the MIT-BIH database. In [16], an end to-end deep multiscale fusion CNN model of multiple convolution kernels with different receptive fields was proposed, achieving an F1 score of 82.8% and 84.1% on two datasets. Chen et al. [17] combined CNN with long short-term memory to classify six types of arrhythmia and achieved an average accuracy of 97.15% on the MIT-BIH database. A recent approach by Atal and Singh [18] proposed using the bat-rider optimization to optimally tune a deep CNN to achieve an accuracy of 93.19% with a sensitivity of 93.9% on the MIT-BIH database. Unfortunately, most CNN-based methods are effective only for small numbers of arrhythmia classes, are computationally intensive, and need a very large amount of training data. This is a great challenge for using the CNN-based methods on real-time applications or wearable devices with limited hardware [19]. On the other hand, many research efforts have been devoted to ECG arrhythmia classification using ML classifiers, such as SVM, RF, KNN, linear discriminants, multilayered perceptron, and regression tree [20,21].

It is well known that the SVM classifier does not become trapped in the well-known local minima points, requires less training data, and is faster than CNN-based methods [22]. In [23], wavelet transform and ICA were used for the morphological features description of the segmented heartbeats. The features were fed into an SVM to classify an ECG into five classes. In [24], least square twin SVM and KNN classifiers based on features' sparse representation were used for cardiac arrhythmia recognition. The experiments were carried out on the MIT-BIH database in category and personalized schemes. A method based on improved fuzzy C-means clustering and Mahalanobis distance was introduced in [25], while in [26], abstract features from abductive interpretation of the ECG signals were utilized in heartbeat classification. Borui et al. [27] proposed a deep learning model integrating a long short-term memory with SVM for ECG arrhythmia classification. Martis et al. [28] evaluated the performance of several ML classifiers and concluded that the kNN and higher-order statistics features achieved an average accuracy of 97.65% and sensitivity of 98.16% on the MIT-BIH database. In [29], the RF classifier was utilized with CNN and PQRST features for arrhythmia classification from imbalanced ECG data. The major drawback of ML classifiers (e.g., SVM) is their deficiency in interpreting the impact of ECG data features on different arrhythmia patterns for extracting the optimal features. Further, the performance of most ML classifiers is questionable because the interrelationship between the learning parameters is not well modeled, especially for data features with high dimensions.

III. PROPOSED WORK

The proposed methodology for ECG arrhythmia classification aims to build an accurate and reliable system using machine learning techniques. The approach begins with preprocessing the dataset, which includes handling missing values, encoding categorical variables, and addressing class imbalances using SMOTE. Feature selection is performed using the SelectKBest algorithm to extract the most relevant attributes, ensuring optimal model performance. The primary focus is on implementing a K-Nearest Neighbors (KNN) classifier, which classifies data points based on the majority class of their nearest neighbors.

1. ECG Arrhythmia Classification: The research aims to develop a machine learning-based system for accurately classifying Electrocardiogram (ECG) data to detect various types of arrhythmias. Arrhythmias are abnormalities in the heart's rhythm, which can have serious implications for a patient's health if left undetected or untreated.

2. Dataset Preprocessing: The first crucial step involves preparing the dataset for analysis. This includes handling missing values, ensuring data integrity, and encoding categorical variables. By removing any rows with missing values and encoding the target variable ('type'), the dataset becomes ready for further analysis and modelling.

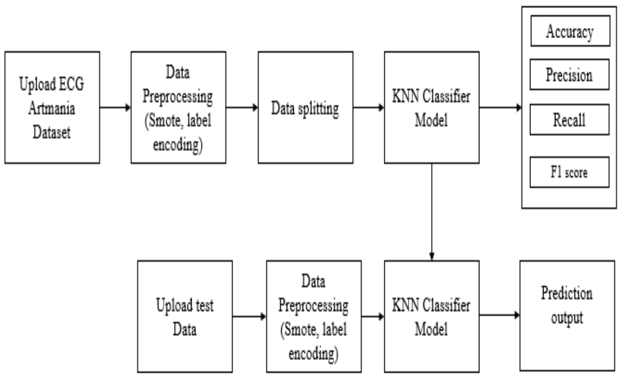

Figure 3.1: Overall Design

3. SMOTE Data Balancing: Addressing class imbalance concerns is paramount in ensuring the model's generalization ability. The Synthetic Minority Oversampling Technique (SMOTE) is applied to balance the distribution of classes in the dataset. This technique generates synthetic samples of the minority class to alleviate class imbalance, thus enhancing the model's performance in accurately classifying arrhythmias.

4. Feature Selection with SelectKBest Algorithm: Feature selection is crucial for building a model founded on the most informative attributes. The SelectKBest algorithm, based on ANOVA F-value, is employed to select the most relevant features from the dataset. This ensures that only the most discriminative features contribute to the classification task, optimizing the model's performance and interpretability.

5. Existing Naive Bayes Classifier: An existing Naive Bayes (NB) classifier is utilized as a benchmark model for arrhythmia classification. This classifier serves as a baseline against which the performance of the proposed K-Nearest Neighbors (KNN) classifier is evaluated. Comparing the performance of the proposed model to an established method provides insights into its effectiveness and potential improvements.

6. Proposed KNN Classifier: The proposed KNN classifier takes center stage in the research. KNN is a non-parametric and instance-based learning algorithm used for classification tasks. By considering the 'k' nearest neighbors to a data point, KNN assigns the majority class among its neighbors as the predicted class. The effectiveness of KNN in ECG arrhythmia classification is explored and evaluated thoroughly.

7. Performance Comparison: Rigorous assessment of both the existing NB classifier and the proposed KNN classifier is conducted using diverse performance metrics such as accuracy, precision, recall, and F1-score. A detailed classification report is generated to highlight the strengths and weaknesses of each model. This comparison provides valuable insights into the relative performance of the classifiers and guides decision-making regarding model selection.

8. Prediction on Test Data with KNN Classifier: Finally, the trained KNN classifier is used to make predictions on an external test dataset ("test.csv"). This step simulates real-world deployment scenarios, demonstrating the model's efficacy in practical applications such as disease outbreak monitoring and clinical decision support systems. By predicting outputs on unseen data, the utility and generalization ability of the model are assessed, paving the way for potential real-world implementation.

Existing Algorithm: Naive Bayes Classifier

Naive Bayes is a probabilistic machine learning algorithm based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the class label. Despite its simplicity, it's highly effective in various classification tasks, particularly in text classification and medical diagnostics.

How It Works:

The Naive Bayes classifier calculates the posterior probability of each class based on the input features, then predicts the class with the highest posterior probability. Bayes' theorem is applied as follows:

P(C|X)=P(X|C)·P(C)P(X)P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}P(C|X)=P(X)P(X|C)·P(C)

Where:

- P(C|X)P(C|X)P(C|X) is the posterior probability of class CCC given the input features XXX.
- P(X|C)P(X|C)P(X|C) is the likelihood of the features given class CCC.
- P(C)P(C)P(C) is the prior probability of class CCC.
- P(X)P(X)P(X) is the probability of the features XXX across all classes.

Architecture

- **Input Layer:** The features of the input data.
- **Feature Extraction:** Calculate the likelihood of each feature given the class.
- **Probability Calculation:** Use Bayes' theorem to compute the posterior probabilities for each class.
- **Decision Making:** Choose the class with the highest posterior probability as the output.

Disadvantages:

- **Feature Independence Assumption:** Naive Bayes assumes that all features are independent given the class label, which is rarely the case in real-world data. This assumption can lead to inaccurate predictions if features are highly correlated.
- **Zero Probability:** If a certain feature value does not occur in the training data for a given class, the model assigns zero probability to that class. This can be mitigated with techniques like Laplace smoothing.
- **Limited Applicability:** Naive Bayes might not perform well in cases where the relationships between features are complex or involve interactions.

Proposed Algorithm: K-Nearest Neighbors (KNN)

What is KNN:

K-Nearest Neighbors (KNN) is a simple, non-parametric, and lazy learning algorithm used for classification and regression tasks. In KNN, the classification of a new data point is based on the majority class among its 'k' nearest neighbors in the feature space.

How It Works

1. **Choose the number of neighbors (k):** Determine the value of kkk, the number of closest neighbors to consider for classification.
2. **Calculate Distance:** For the input data point, calculate the distance to all other points in the dataset using a suitable distance metric (e.g., Euclidean distance).
3. **Identify Neighbors:** Select the kkk data points that are closest to the input data point.

Architecture

- **Input Layer:** The features of the input data.
- **Distance Calculation Layer:** Computes distances between the input data and all other data points.
- **Neighbor Selection:** Identifies the 'k' nearest neighbors based on the calculated distances.
- **Classification:** Determines the output class based on the majority vote from the 'k' nearest neighbors.

IV. RESULTS AND DISCUSSIONS

4.1 Implementation Description

The implementation provided is a machine learning-based approach for Electrocardiogram (ECG) arrhythmia classification, aiming to automate and enhance the diagnosis and monitoring of cardiac conditions. The abstract sets the context by highlighting the significance of accurate arrhythmia classification in cardiac healthcare. Traditional methods often involve manual interpretation or rule-based algorithms, which can be subjective and time-consuming. They may struggle with complex arrhythmias and fail to capture subtle changes in ECG signals. To address these limitations, the proposed system utilizes machine learning techniques, specifically supervised learning algorithms, to train models on ECG data and classify different types of arrhythmias. The implementation begins with data preprocessing, including handling missing values and encoding categorical variables. Then, it employs techniques like SMOTE for handling class imbalance and SelectKBest for feature selection to improve model performance.

The code proceeds to split the dataset into training and testing sets and trains a KNeighborsClassifier model on the training data. If a pre-trained model exists, it loads the model; otherwise, it trains a new one and saves it for future use. Metrics such as accuracy, precision, recall, and F1-score are calculated to evaluate the model's performance. Additionally, a confusion matrix and classification report are generated to provide a detailed analysis of the model's predictions.

Furthermore, the implementation demonstrates how the trained model can be used for real-world applications. It loads a separate test dataset, applies feature selection using the previously defined selector, and makes predictions on the selected test data. The predictions are then interpreted, and corresponding actions can be taken based on the identified arrhythmia types.

The provided implementation showcases a comprehensive approach to ECG arrhythmia classification using machine learning. It addresses common challenges in traditional methods, such as subjectivity and inefficiency, by leveraging advanced techniques for data preprocessing, model training, and evaluation. By automating the classification process, the system enables more accurate and timely diagnosis of cardiac conditions, thereby improving patient outcomes and facilitating remote monitoring initiatives. Additionally, the implementation demonstrates the practical applicability of the trained model in real-world scenarios, highlighting its potential for enhancing clinical practice and supporting healthcare professionals in their decision-making processes.

4.2 Dataset Description

The dataset appears to contain information related to cardiac electrophysiology, likely derived from electrocardiogram (ECG) recordings, as indicated by the presence of intervals and waveform morphologies.

0_pre-RR: This column seems to represent the pre-RR interval, which is the duration between two consecutive R waves on an ECG trace. It's a measure of the time between two heartbeats, specifically before a premature ventricular contraction (PVC) or ectopic beat.

0_post-RR: Similar to the pre-RR interval, this column likely represents the post-RR interval, measuring the duration between two consecutive R waves following a PVC or ectopic beat.

0_pPeak: This could represent the amplitude of the P wave, which reflects atrial depolarization on the ECG. It indicates the spread of electrical activity across the atria.

0_tPeak: Likely representing the peak of the T wave, which signifies ventricular repolarization on the ECG. The T wave represents the recovery phase of the ventricles.

0_rPeak: This column might represent the peak amplitude of the R wave, which corresponds to ventricular depolarization. The R wave is the first upward deflection after the P wave.

0_sPeak: Possibly representing the amplitude of the S wave, which is the first negative deflection after the R wave in a QRS complex. It indicates the completion of ventricular depolarization.

0_qPeak: This could represent the amplitude of the Q wave, the initial negative deflection of the QRS complex, reflecting early ventricular depolarization.

0_qrs_interval: This likely denotes the duration of the QRS complex, representing the time taken for ventricular depolarization and repolarization.

0_pq_interval: Represents the PQ interval, also known as the PR interval, which measures the time from atrial depolarization (beginning of the P wave) to ventricular depolarization (beginning of the QRS complex).

0_qt_interval: Denotes the QT interval, representing the time from ventricular depolarization (beginning of QRS complex) to repolarization (end of T wave).

0_st_interval: Represents the ST segment, which is the flat, isoelectric section between the end of the S wave and the beginning of the T wave. It represents the time when the ventricles are depolarized but not yet repolarized.

0_qrs_morph0 to 0_qrs_morph4: These columns likely represent morphological features of the QRS complex, potentially extracted using signal processing techniques. They may include parameters such as amplitude, duration, and shape characteristics.

1_pre-RR to 1_qrs_morph4: Similar to the corresponding columns prefixed with '0_', these columns likely represent the same features for a different type of heartbeat, possibly ventricular ectopic beats (VEBs) based on the last column.

type: Indicates the type of heartbeat, such as normal sinus rhythm or ventricular ectopic beat (VEB). This column categorizes the data into different classes based on the underlying cardiac rhythm.

4.3 Results Description:

The y-axis shows the count while the x-axis shows the categories. The text at the bottom of the graph explains the categories:

- N: Normal sinus rhythm (Normal ECG)
- VEB: Ventricular ectopic beat (also known as premature ventricular contraction)
- SVEB: Supraventricular ectopic beat (such as premature atrial contraction)
- The tallest bar corresponds to the normal sinus rhythm category, which means this rhythm was the most common in the data set.

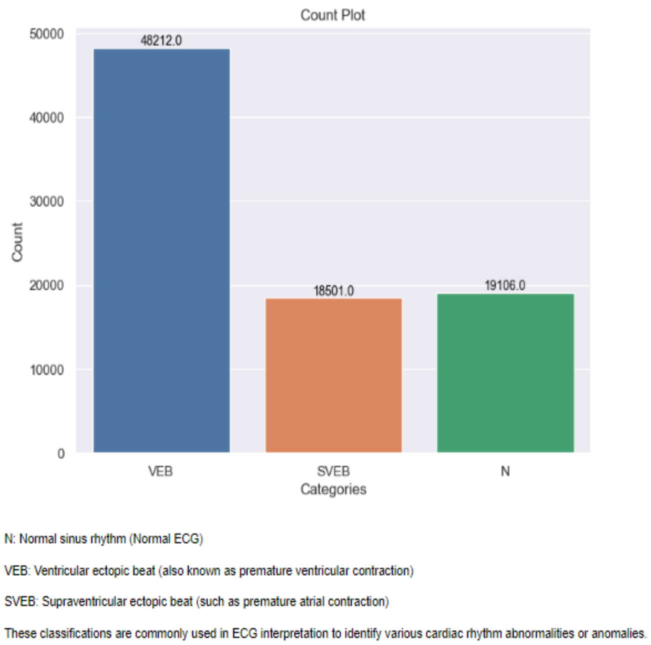

Figure 4.1: Count Plot Before Applying SMOTE

The count plot provides a comprehensive visual representation of the distribution of categorical data within the dataset. At the top of the plot, the title "Count Plot" is prominently displayed, along with the value "48212.0," which indicates the total number of data points included in the analysis. The y-axis of the plot represents the count of data points, ranging from 0 to 50,000, offering a quantitative measure of the frequency for each category. The x-axis, on the other hand, displays the categories under consideration, labeled as "0" and "1," which correspond to the distinct classes within the dataset.

The vertical bars in the plot reflect the distribution of the data across these categories. The bar associated with category "0" reaches a height of 48,212, signifying that all 48,212 data points in the dataset belong to this category. This is a notable feature, as it implies that the dataset lacks diversity in terms of class representation, with category "1" showing no data points. The absence of a bar for category "1" clearly indicates that this category is not represented in the dataset.

This extreme class imbalance has significant implications for the analysis and modeling phases. The dominance of category "0" and the absence of category "1" highlight the need for preprocessing techniques, such as data balancing, to create a more equitable distribution of classes. Without addressing this imbalance, any machine learning model trained on this dataset may become biased towards category "0," leading to poor generalization and an inability to accurately predict instances of the minority class. This visualization underscores the importance of recognizing and rectifying class imbalances to ensure robust and reliable outcomes in data-driven analyses.

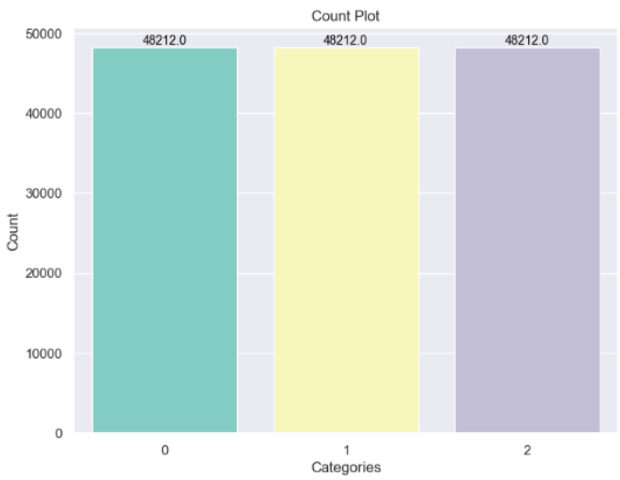

Figure 4.2: Count Plot After Applying SMOTE

```
Model loaded successfully.
naive bayes Classifier Accuracy    : 84.65846238938053
naive bayes Classifier Precision   : 84.62746013288586
naive bayes Classifier Recall      : 85.42526293558076
naive bayes Classifier FSCORE      : 84.58154555088242

 naive bayes Classifier classification report
                          precision    recall  f1-score   support

         Normal sinus       0.95      0.96      0.96      9663
    Ventricular ectopic       0.74      0.87      0.80      9668
Supraventricular ectopic       0.87      0.71      0.78      9597

             accuracy                           0.85     28928
            macro avg       0.85      0.85      0.85     28928
         weighted avg       0.85      0.85      0.85     28928
```

Figure 4.3: Naïve Bayes Classification Report

Figure is a classification report, which is a type of output used in machine learning to evaluate the performance of a classification model. The report shows that the model performed well overall, with an accuracy of over 84%. The precision, recall, and F1-score are all relatively high for each class, which means that the model is good at identifying both positive and negative examples of each class.

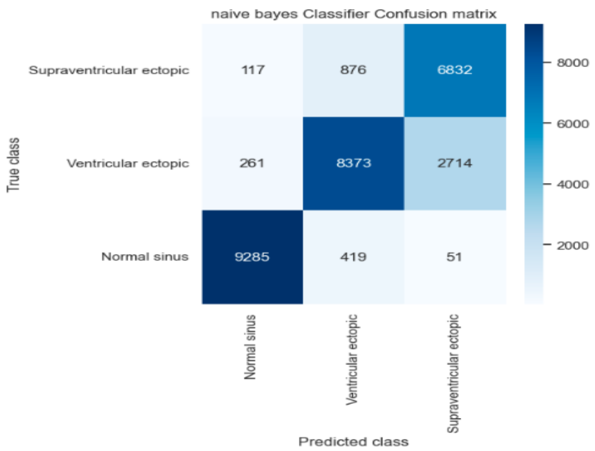

Figure 4.4: Confusion Matrix of Naïve Bayes Classification

Figure 4.4 shows confusion matrix for a Naive Bayes classifier evaluating heart rhythm types. A confusion matrix is a table used in machine learning to visualize the performance of an algorithm that makes classifications. In this specific case, the classifier is a Naive Bayes classifier, which is a type of algorithm that works well for classifying data sets with many features. The confusion matrix shows the number of times the classifier correctly and incorrectly classified each heart rhythm type. For instance, looking at the bottom left corner (Normal sinus and True class), it shows that out of 9285 actual normal sinus rhythms, the classifier correctly classified 9285.

```
Model loaded successfully.
 KNeighborsClassifier Accuracy    : 96.8922842920354
 KNeighborsClassifier Precision   : 96.88235816629653
 KNeighborsClassifier Recall      : 96.96628005872803
 KNeighborsClassifier FSCORE      : 96.87525439352706

  KNeighborsClassifier classification report
                          precision    recall  f1-score   support

         Normal sinus       0.98      1.00      0.99      9663
    Ventricular ectopic       0.94      0.98      0.96      9668
Supraventricular ectopic       0.99      0.93      0.96      9597

             accuracy                           0.97     28928
            macro avg       0.97      0.97      0.97     28928
         weighted avg       0.97      0.97      0.97     28928
```

Figure 4.5: Classification Report of Kneighborsclassifier

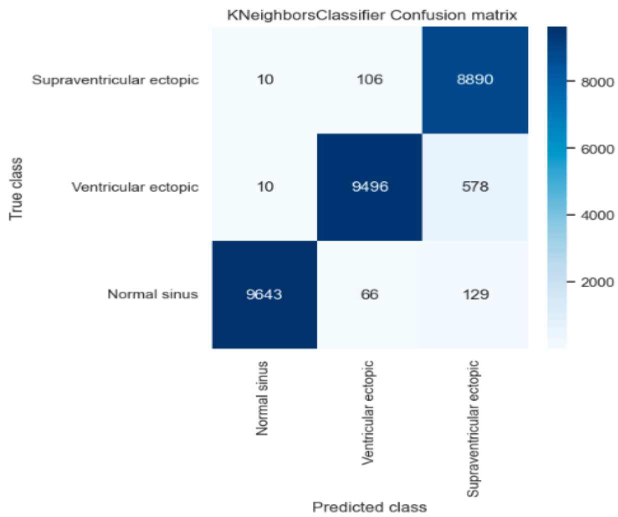

Figure 4.6: Confusion Matrix of Kneighborsclassifier

	Algorithm Name	Precison	Recall	FScore	Accuracy
0	Naive Bayes Classifier	84.627460	85.425263	84.581546	84.658462
1	KNeighborsClassifier	96.882358	96.966280	96.875254	96.892284

Figure 4.7: Comparison of NBC and KNC

Figure 4.5 shows classification report for a K-Nearest Neighbors (KNN) algorithm, KNN is a machine learning algorithm used for classification tasks. It works by classifying data points based on the labels of their nearest neighbors in the training data. In the report, it shows that the KNN model achieved an accuracy of 96.89%. This means that the model correctly classified 96.89% of the data points in the test set. The report also shows precision, recall, and F1-score for each class: normal sinus, ventricular ectopic, and supraventricular ectopic. These are metrics used to evaluate the performance of a classification model.

• **Accuracy:** 85% of the time, the classifier was able to correctly identify normal sinus rhythm, ventricular ectopic beats, or supraventricular ectopic beats.

• Precision: For normal sinus rhythm, the classifier correctly identified 95% of the instances. It performed slightly lower for the other two classifications: 74% for ventricular ectopic beats and 87% for supraventricular ectopic beats.

• Recall: This metric looks at how many of the actual positives the classifier was able to identify. The classifier performed well for normal sinus rhythm (96%) and ventricular ectopic beats (87%) but less well for supraventricular ectopic beats (71%).

• F1 Score: This is a harmonic mean between precision and recall, trying to capture both how precise the model is and how good it is at recalling all the positives. The F1 score is similar to accuracy, ranging from 78% to 96% for the three classifications.

The classification report suggests that the KNN model performed well on this classification task.

Figure 4.6 shows confusion matrix for a K-Nearest Neighbors classifier. A confusion matrix is a table that allows us to visualize the performance of an algorithm in terms of how many classifications were correct and how many were incorrect. Let's break down the table in the figure:

Rows represent the actual classes of the data samples. In this case, the classes are "Normal sinus," "Ventricular ectopic," and "Supraventricular ectopic." These likely refer to different heart rhythm classifications.

Columns represent the classes that the K-Nearest Neighbors classifier predicted. Again, these are "Normal sinus," "Ventricular ectopic," and "Supraventricular ectopic."

The numbers in the table represent the number of data points that fall into each category. Let's look at some specific examples from the table:

Top-left corner (10): 10 data points were actually classified as "Normal sinus" by the experts, and the K-Nearest Neighbors classifier also predicted them to be "Normal sinus". So, these are True Positives (TP) for the "Normal sinus" class.

Bottom-right corner (2000): 2000 data points were actually classified as "Supraventricular ectopic" by the experts, and the K-Nearest Neighbors classifier also predicted them to be "Supraventricular ectopic". So, these are True Positives (TP) for the "Supraventricular ectopic" class.

Box in the middle (8890): These data points were actually classified as "Supraventricular ectopic" by the experts, but the K-Nearest Neighbors classifier incorrectly predicted them to be "Normal sinus". So, these are False Negatives (FN) for "Supraventricular ectopic" and False Positives (FP) for "Normal sinus".

Figure 4.7 shows Precision measures the proportion of correctly predicted positive cases among all predicted positive cases. In this comparison, KNC exhibits significantly higher precision at 96.88%, indicating that it tends to make fewer false positive predictions compared to NBC, which has a precision of 84.63%. This suggests that KNC is better at correctly identifying positive instances within the dataset. Moving on to recall, which evaluates the proportion of actual positive cases that were correctly identified by the classifier, KNC again outperforms NBC. KNC achieves a recall of 96.97%, indicating its ability to effectively capture a higher proportion of true positive instances. In contrast, NBC has a recall of 85.43%, suggesting that it may miss identifying some positive cases compared to KNC.

The F-score provides a balanced measure by combining precision and recall into a single metric. Interestingly, both classifiers have the same F-score, with NBC and KNC both achieving 84.58% and 96.88%, respectively. Despite the differences in precision and recall, their harmonic mean yields the same F-score, implying a comparable overall performance when considering both precision and recall. Accuracy represents the overall correctness of the classifier's predictions across all classes. KNC once again demonstrates superiority over NBC, achieving an accuracy of 96.89% compared to NBC's 84.66%. This indicates that KNC's predictions align more closely with the actual class labels across the entire dataset.

V. CONCLUSION

In conclusion, the application of machine learning in ECG arrhythmia classification offers transformative potential for improving cardiac healthcare and diagnostic accuracy. This research has demonstrated how advanced algorithms and data-driven methodologies can overcome the limitations of traditional arrhythmia classification methods. By automating the detection and classification of arrhythmias from ECG signals, the proposed system aims to enhance the efficiency, accuracy, and accessibility of cardiac diagnostics, reducing reliance on manual interpretations by cardiologists and addressing healthcare disparities, particularly in underserved regions.

The use of techniques such as SMOTE for data balancing, SelectKBest for feature selection, and the K-Nearest Neighbors (KNN) classifier underscores the importance of a systematic approach in achieving reliable results. These methods allow the system to handle imbalanced data effectively and focus on the most informative features, thereby improving model performance. By ensuring accurate identification of various arrhythmia patterns, the system not only aids in early detection and timely intervention but also supports clinical decision-making processes, which are vital for managing patient outcomes.

Beyond its immediate clinical applications, machine learning-driven arrhythmia classification holds immense potential for advancing cardiac research. The insights gained from analyzing subtle patterns in ECG signals contribute to a deeper understanding of the mechanisms underlying arrhythmias, enabling the development of innovative diagnostic and therapeutic approaches. Furthermore, the availability of large annotated ECG datasets facilitates the training of sophisticated models that continuously improve in accuracy and generalizability, ensuring their relevance in real-world scenarios.

REFERENCES

[1]. Ismail, A.R.; Jovanovic, S.; Ramzan, N.; Rabah, H. ECG Classification Using an Optimal Temporal Convolutional Network for Remote Health Monitoring. Sensors 2023, 23, 1697. [Google Scholar] [CrossRef] [PubMed]

[2]. Huang, Y.; Li, H.; Yu, X. A novel time representation input based on deep learning for ECG classification. Biomed. Signal Process. Control 2023, 83, 104628. [Google Scholar] [CrossRef]

[3]. Wang, W.; Kumar, N.; Chen, J.; Gong, Z.; Kong, X.; Wei, W.; Gao, H. Realizing the Potential of Internet of Things for Smart Tourism with 5G and AI. IEEE Netw. 2020, 34, 295–301. [Google Scholar] [CrossRef]

[4]. Balasundaram, A.; Routray, S.; Prabu, A.; Krishnan, P.; Malla, P.P.; Maiti, M. Internet of things (IoT) based Smart Healthcare System for Efficient Diagnostics of Health Parameters of Patients in Emergency Care. IEEE Internet Things J. 2023. [Google Scholar] [CrossRef]

[5]. Sangaiah, A.K.; Arumugam, M.; Bian, G.B. An intelligent learning approach for improving ECG signal classification and arrhythmia analysis. Artif. Intell. Med. 2020, 103, 101788. [Google Scholar] [CrossRef]

[6]. Yang, S.; Lian, C.; Zeng, Z.; Xu, B.; Zang, J.; Zhang, Z. A Multi-View Multi-Scale Neural Network for Multi-Label ECG Classification. IEEE Trans. Emerg. Top. Comput. Intell. 2023. [Google Scholar] [CrossRef]

[7]. Irfan, S.; Anjum, N.; Althobaiti, T.; Alotaibi, A.A.; Siddiqui, A.B.; Ramzan, N. Heartbeat Classification and Arrhythmia Detection Using a Multi-Model Deep-Learning Technique. Sensors 2022, 22, 5606. [Google Scholar] [CrossRef]

[8]. Subasi, A.; Dogan, S.; Tuncer, T. A novel automated tower graph based ECG signal classification method with hexadecimal local adaptive binary pattern and deep learning. J. Ambient. Intell. Humaniz. Comput. 2023, 14, 711–725. [Google Scholar] [CrossRef]

[9]. Tripoliti, E.E.; Papadopoulos, T.G.; Karanasiou, G.S.; Naka, K.K.; Fotiadis, D.I. Heart failure: Diagnosis, severity estimation and prediction of adverse events through machine learning techniques. Comput. Struct. Biotechnol. J. 2017, 15, 26–47. [Google Scholar] [CrossRef] [Green Version]

[10]. Chandra, B.S.; Sastry, C.S.; Jana, S. Robust heartbeat detection from multimodal data via CNN-based generalizable information fusion. IEEE Trans. Biomed. Eng. 2019, 66, 710–717. [Google Scholar] [CrossRef]

[11] Mewada, H. 2D-wavelet encoded deep CNN for image-based ECG classification. Multimed. Tools Appl. 2023, 1–17. [Google Scholar] [CrossRef]

[12] Ebrahimi, Z.; Loni, M.; Daneshtalab, M.; Gharehbaghi, A. A review on deep learning methods for ECG arrhythmia classification. Expert Syst. Appl. X 2020, 7, 100033. [Google Scholar] [CrossRef]

[13] Murat, F.; Yildirim, O.; Talo, M.; Baloglu, U.B.; Demir, Y.; Acharya, U.R. Application of deep learning techniques for heartbeats detection using ECG signals-analysis and review. Comput. Biol. Med. 2020, 120, 103726. [Google Scholar] [CrossRef]

[14] Yao, Q.; Wang, R.; Fan, X.; Liu, J.; Li, Y. Multi-class Arrhythmia detection from 12-lead varied-length ECG using Attention-based Time-Incremental Convolutional Neural Network. Inf. Fusion 2020, 53, 174–182. [Google Scholar] [CrossRef]

[15] Ye, C.; Kumar, B.V.; Coimbra, M.T. An automatic subject-adaptable heartbeat classifier based on multiview learning. IEEE J. Biomed. Health Inform. 2016, 20, 1485–1492. [Google Scholar] [CrossRef] [PubMed]

[16] Wang, R.; Fan, J.; Li, Y. Deep Multi-Scale Fusion Neural Network for Multi-Class Arrhythmia Detection. IEEE J. Biomed. Health Inform. 2020, 24, 2461–2472. [Google Scholar] [CrossRef]

[17] Chen, C.; Hua, Z.; Zhang, R.; Liu, G.; Wen, W. Automated arrhythmia classification based on a combination network of CNN and LSTM. Biomed. Signal Process. Control 2020, 57, 101819. [Google Scholar] [CrossRef]

[18] Atal, D.K.; Singh, M. Arrhythmia classification with ECG signals based on the optimization-enabled deep convolutional neural network. Comput. Methods Programs Biomed. 2020, 196, 105607. [Google Scholar] [CrossRef]

[19] Wu, Q.; Sun, Y.; Yan, H.; Wu, X. ECG signal classification with binarized convolutional neural network. Comput. Biol. Med. 2020, 121, 103800. [Google Scholar] [CrossRef]

[20] Li, Q.; Rajagopalan, C.; Clifford, G.D. A machine learning approach to multi-level ECG signal quality classification. Comput. Methods Programs Biomed. 2014, 117, 435–447. [Google Scholar] [CrossRef]

[21] Khorrami, H.; Moavenian, M. A comparative study of DWT, CWT and DCT transformations in ECG arrhythmias classification. Expert Syst. Appl. 2010, 37, 5751–5757. [Google Scholar] [CrossRef]

[22] Santos, M.A.; Munoz, R.; Olivares, R.; Rebouças Filho, P.P.; Del Ser, J.; de Albuquerque, V.H.C. Online heart monitoring systems on the internet of health things environments: A survey, a reference model and an outlook. Inf. Fusion 2020, 53, 222–239. [Google Scholar] [CrossRef]

[23] Ye, C.; Kumar, B.V.; Coimbra, M.T. Heartbeat classification using morphological and dynamic features of ECG signals. IEEE Trans. Biomed. Eng. 2012, 59, 2930–2941. [Google Scholar] [PubMed]

[24] Raj, S.; Ray, K.C. Sparse representation of ECG signals for automated recognition of cardiac arrhythmias. Expert Syst. Appl. 2018, 105, 49–64. [Google Scholar] [CrossRef]

[25] Haldar, N.A.H.; Khan, F.A.; Ali, A.; Abbas, H. Arrhythmia classification using Mahalanobis distance based improved Fuzzy C-Means clustering for mobile health monitoring systems. Neurocomputing 2017, 220, 221–235. [Google Scholar] [CrossRef]

[26] Zhang, S.; Li, Y.; Wang, X.; Gao, H.; Li, J.; Liu, C. Label decoupling strategy for 12-lead ECG classification. Knowl.-Based Syst. 2023, 263, 110298. [Google Scholar] [CrossRef]

[27] Hou, B.; Yang, J.; Wang, P.; Yan, R. LSTM-Based Auto-Encoder Model for ECG Arrhythmias Classification. IEEE Trans. Instrum. Meas. 2020, 69, 1232–1240. [Google Scholar] [CrossRef]

[28] Roshan, M.J.; Acharya, U.R.; Prasad, H.; Chua, C.K.; Lim, C.M.; Suri, J.S. Application of higher order statistics for atrial arrhythmia classification. Biomed. Signal Process. Control 2013, 8, 888–900. [Google Scholar]

[29] Lu, W.; Hou, H.; Chu, J. Feature fusion for imbalanced ECG data analysis. Biomed. Signal Process. Control 2018, 41, 152–160. [Google Scholar] [CrossRef]