LARGE VOLUME ECG SENSOR DATA CLASSIFICATION AND ASSOCIATION RULES

Otabek Kadamboyevich Khujaev
Azizbek Dilshodovich Jumanazarov
Urgench Branch of Tashkent University of
Information Technologies named after Muhammad al-Khwarizmi
otabek.hujaev@gmail.com, devdilshodovich@gmail.com

ABSTRACT

This paper explores the classification of large volumes of electrocardiogram (ECG) sensor data using machine learning techniques. The aim is to develop an accurate and efficient system for categorizing ECG signals into different classes based on their features. Furthermore, the study investigates the use of association rules to uncover patterns and relationships between different ECG classes. The proposed system utilizes various algorithms and techniques, including decision trees, support vector machines, and random forests, to classify ECG data. The results indicate that the proposed system achieves high accuracy and can effectively classify large volumes of ECG data. Additionally, the use of association rules provides valuable insights into the relationships between different ECG classes, which can aid in the diagnosis and treatment of cardiovascular diseases.

Keywords: Association Rule, ECG, CVD, Classification, Deep Learning, Health, MIT-BIH database.

INTRODUCTION

An Electrocardiogram (ECG) is a medical test that records the electrical activity of the heart over a period of time. ECG sensor data is widely used in clinical practice and research, and is an important tool for diagnosing and monitoring a variety of heart conditions, including arrhythmias, myocardial infarction, and heart failure. In 2020, approximately 19.1 million deaths were attributed to cardiovascular disease (CVD) globally. The age-adjusted death rate per 100,000 population was 239.8. The age-adjusted prevalence rate was 7354.1 per 100,000. The mortality rates as a result of CVD were the highest in Eastern Europe and Central Asia in the year 2020. Several other regions, including Oceania, North Africa, the Middle East, Central Europe, sub-Saharan Africa, and South and Southeast Asia, also experienced relatively high mortality rates due to CVD. Conversely, regions such as high-income Asia Pacific and North America, Latin America, Western Europe, and Australasia had the lowest rates of mortality [1].

The aim of this research is to explore the association rules within large volumes of ECG sensor data, which requires the classification of the data. After classification, researchers can explore the patterns and relationships between variables to identify significant correlations or dependencies. This exploration could provide valuable insights for medical research and diagnosis.

In Section 2, a review of relevant works is presented, whereas Section 3 outlines the method proposed in this study. Finally, Section 4 provides the conclusion.

METHODS

In a study by Themis P. Exarchos [2] a new methodology was introduced for the automated detection of ischemic beats, utilizing classification through association rules. The proposed methodology offers the advantage of high accuracy combined with the ability to explain the decisions made through the use of association rules. The results of the study demonstrate the effectiveness of the approach in comparison to previous studies using the same subset from the ESC ST-T [3] database, suggesting that it could be integrated into a system for detecting ischemic episodes in long ECG recordings. However, further evaluation through clinical testing is required to fully assess its potential.

Tanis Mar [4]. This study explores the use of a suitable feature selection (FS) procedure to improve the performance of ECG classifiers while reducing their complexity, which can be highly beneficial for online ECG monitoring in ambulatory settings. A new performance measure index was introduced to address class imbalance and the relative importance of different arrhythmias in heartbeat classification. The algorithm was executed on two sets of features, with the second set focused specifically on identifying features suitable for online monitoring. The results of the study demonstrate the effectiveness of the FS procedure in improving classifier performance while reducing complexity. Additionally, the study found that the MLP classifier outperformed linear classifiers in the field of heartbeat classification.

The study of Muhammad Zubair, Jinsul Kim and Changwoo Yoon [5] introduces an ECG heart beat classifier that uses convolutional neural networks to extract and learn appropriate features from raw ECG data. The Massachusetts institute of technology and Beth Israel hospital (MIT-BIH) database was used to evaluate the performance of the proposed ECG beat classification system. The ECG beats were labeled and classified into five beat types according to Association for the Advancement of Medical Instrumentation (AAMI) standards, and a small patient-specific dataset was used for training. The experiment showed that the proposed model achieved significant classification accuracy and excellent computational efficiency. Future work of the team will focus on enhancing performance by comparing the classification accuracy of ventricular and supraventricular beats with other ECG beat classification algorithms using deep learning.

DISCUSSION

The topic of large volume ECG sensor data classification and association rules focuses on the challenge of analyzing and making sense of vast amounts of electrocardiogram (ECG) data generated by sensors. By using classification and association rule techniques, healthcare providers and researchers can identify patterns and relationships in the data, which can improve our understanding of cardiovascular health and lead to more targeted treatment plans for patients. However, there are challenges associated with analyzing large volumes of ECG data, such as noise and variations based on factors like age and gender. Continuing to develop advanced machine learning and data analysis techniques can help overcome these challenges and improve patient outcomes.

Data Collection

This study utilizes a dataset that has been made available by Kaggle. The dataset used in this study comprises two sets of heartbeat signals that are

derived from the MIT-BIH Arrhythmia Dataset and The PTB Diagnostic ECG Database, which are well-known datasets in heartbeat classification. The size of both collections is sufficient for training a deep neural network. The dataset has been used to explore the use of deep neural network architectures for heartbeat classification and to observe the capabilities of transfer learning. The signals in the dataset represent ECG shapes of heartbeats for both normal cases and cases affected by arrhythmias and myocardial infarction. Each signal has been preprocessed and segmented into corresponding heartbeats.

The data consists of 187 columns and contains 109,446 samples that are classified into 5 categories. To work with the data the first step will be to create a pie chart visualizing the distribution of the data in the 187 column of the data frame. After calculating the number of samples in each category the further step will be creating a new figure with a size of 20x10. A circle with a radius of 0.7 and a white color is created and the labels and colors for each pie slice specified using the labels and colors arguments. Then, the percentage of each category displayed on Figure [1]. The pie chart visualizes the distribution of data in the 187 column, where each slice corresponds to a category and its size represents the number of samples in that category. The chart's labels and colors aid in interpreting the data and identifying any imbalances or biases in the dataset.

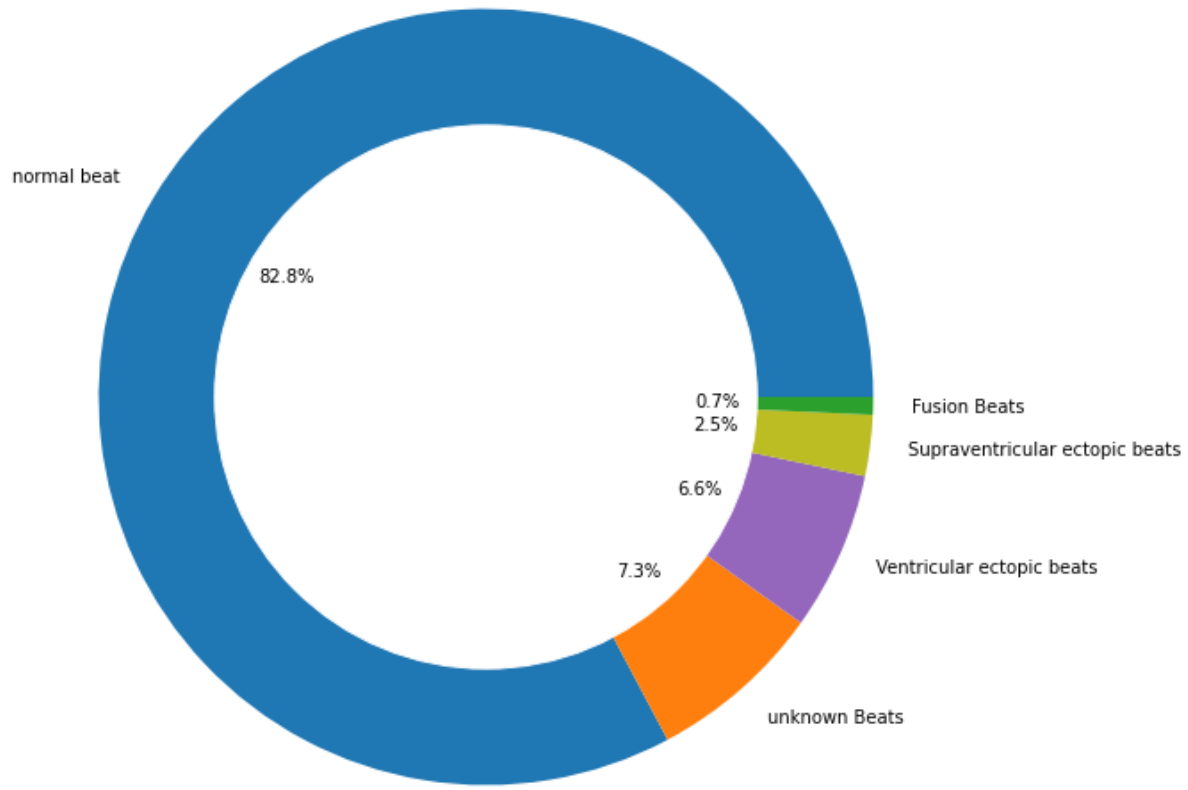

Figure 1. The distribution of data

Resampling for balancing the dataset

In this step we should create five separate data frames based on the different categories in the 187 column. The data frame is also downsampled to 20,000 samples to balance the number of samples across categories. Furthermore, the next step concatenates the downsampled and upsampled data frames into a new data frame that has balanced class representation. This technique of resampling can improve the performance of the machine learning model by preventing it from being biased towards the categories with more samples. After resampling the data frame to balance class representation, it is important to visualize the new class distribution to ensure that it is indeed balanced. This method creates a pie chart Figure [2] to visualize the distribution of the data in the 187 column after resampling. Finally, the method will create a new figure with a size of 20x10. A circle with a radius of 0.7 and a white color is created, then the labels and colors for each pie slice are specified using the labels and colors arguments. The method specifies that the percentage of each category should be displayed.

By visualizing the class distribution, we can ensure that the resampling technique was successful in balancing the number of samples across categories. This can help improve the performance of the machine learning model and prevent it from being biased towards certain categories.

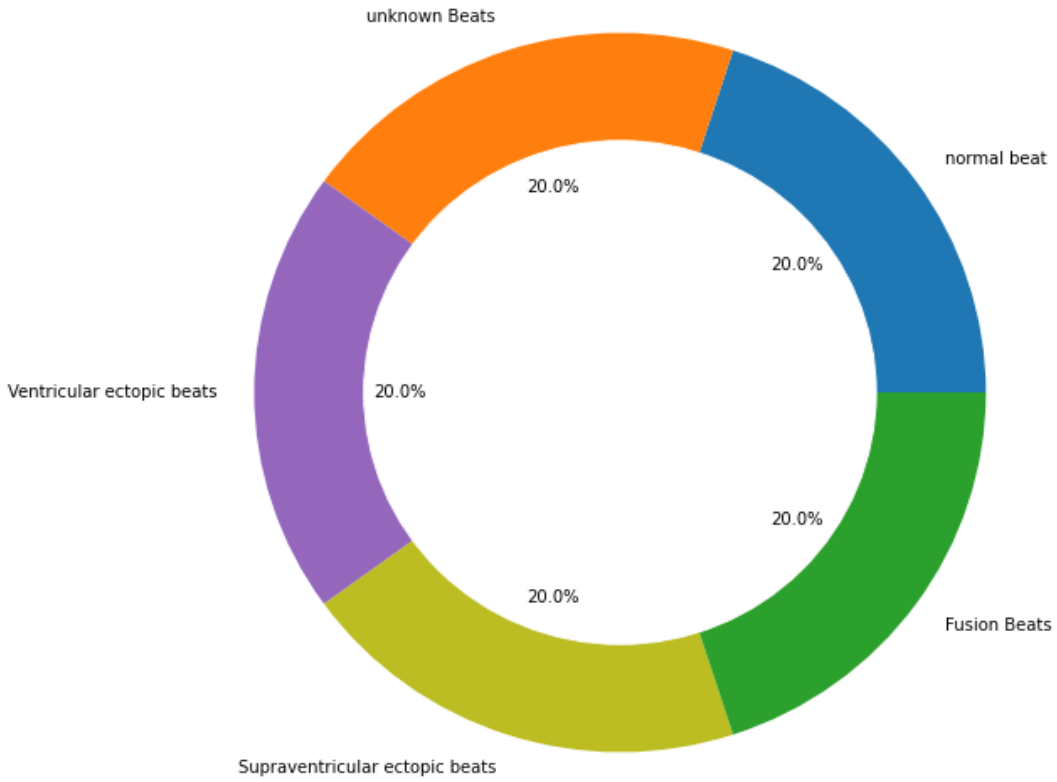

Figure 2. Visualizing the class distribution

Classes

The next technique is frequently employed to generate a smaller subset of the original data frame for either exploratory analysis or to test and

validate the machine learning model. By randomly selecting one sample from each category, we can ensure that the resulting subset is representative of all categories, making it useful for analysis and model validation. Moreover, we create subplots displaying waveform patterns for each category in the 187 column of the data frame. Then the method is used to display the waveform data, and label each subplot. This visualization Figure [3] is helpful for understanding waveform patterns and selecting features for model development in machine learning.

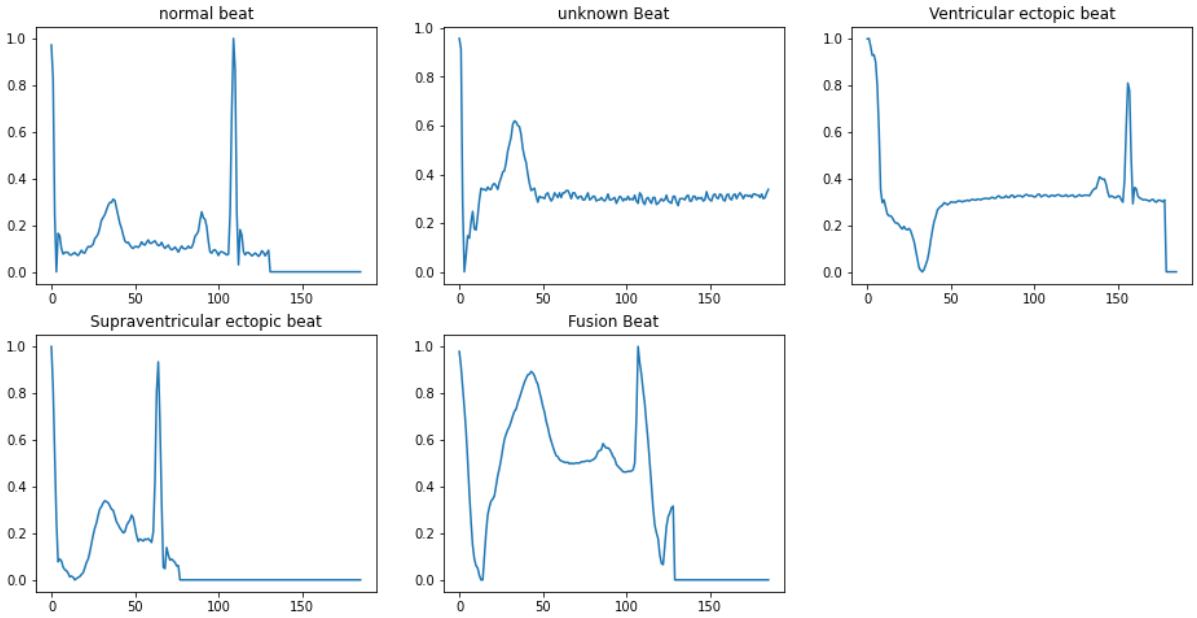

Figure 3. Beat categories

RESULTS

Pretreat

The aim of this study is to develop a machine learning model for classifying electrocardiogram (ECG) signals into five categories: Normal, Unknown, Ventricular Ectopic Beat, Supraventricular Ectopic Beat, and Fusion Beat while considering association rules. The model includes functions for adding Gaussian noise to the ECG signals, defining and training a convolutional neural network (CNN) model on the ECG data, and evaluating the performance of the trained model using metrics such as accuracy and confusion matrix.

The CNN model architecture includes several convolutional layers, max pooling layers, and fully connected layers. The model is trained on the ECG data using the categorical cross-entropy loss and the Adam optimizer. The performance of the trained model is evaluated using accuracy and visualizations of the training and validation loss and accuracy over time Figure [4]. The confusion matrix is also displayed to provide a detailed breakdown of the model's performance across each category. The Accuracy of this model is 98.09%.

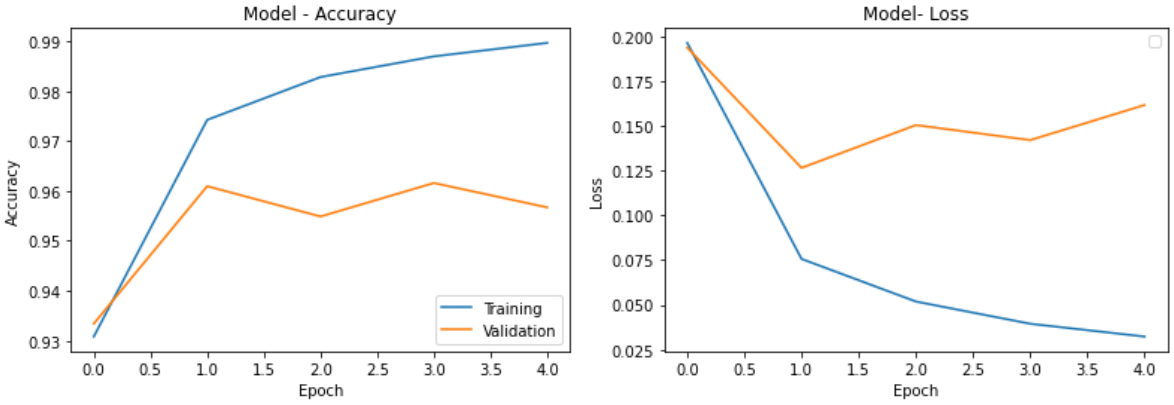

Figure 4. Accuracy and Loss models

The next method defines a function that takes a confusion matrix and class labels as input and plots a visualization of the matrix using matplotlib. The confusion matrix is computed using the function from scikit-learn library, which takes the true labels and predicted labels as input Figure [5].

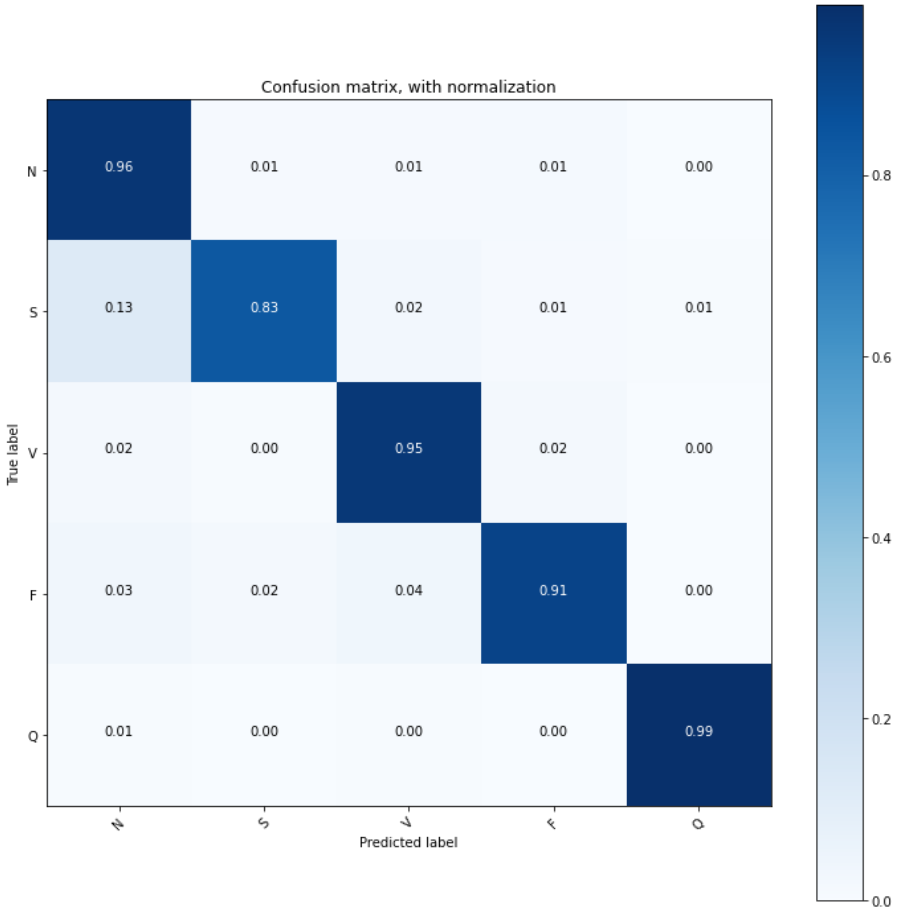

Figure 5. Confusion matrix, with normalization

The function has several optional parameters, including normalize to normalize the confusion matrix, title to set the title of the plot, and set the color map of the plot. The function uses itertools to iterate over the rows and

columns of the confusion matrix and plot the values in each cell. If normalize is set to True, the function normalizes the values in the confusion matrix by dividing each row by its sum. Finally, the method uses plt methods to create a color-coded visualization of the confusion matrix with labeled axes and a color bar. The class labels are also displayed on the x and y axes.

CONCLUSION

In conclusion, Arrhythmia is a common cardiac disorder that can lead to serious health issues if left undiagnosed and untreated. Early and accurate detection is crucial for effective treatment. Our findings uses a CNN model to classify heartbeats into five categories, achieving high accuracy in detecting arrhythmia using the MIT-BIH Arrhythmia Database. The model preprocesses the data by adding Gaussian noise and splits it into training and testing datasets. Overall, our study could help physicians detect arrhythmia more quickly and accurately, improving patient outcomes.

REFERENCES

1. 2022 Heart Disease & Stroke Statistical Update Fact Sheet Global Burden of Disease. American Heart Association, Inc.
2. Themis P. Exarchos, Costas Papaloukas, Dimitrios I. Fotiadis, Lampros K. Michalis. “An Association Rule Mining-Based Methodology for Automated Detection of Ischemic ECG Beats”. IEEE Transactions On Biomedical Engineering, vol. 53, no. 8, August 2006.
3. T. Stamkopoulos, K. Diamantaras, N. Maglaveras, and M. Strintzis, “ECG analysis using nonlinear PCA neural networks for ischemia detection,” IEEE Trans. Signal Process., vol. 46, no. 11, pp. 3058–3067, Nov. 1998.
4. Tanis Mar, Student Member, IEEE, Sebastian Zaunseder, Juan Pablo Martínez, Mariano Llamedo, and Rüdiger Poll. “Optimization of ECG Classification by Means of Feature Selection”. IEEE Transactions On Biomedical Engineering, vol. 58, no. 8, August 2011.
5. Muhammad Zubair, Jinsul Kim, Changwoo Yoon. “An Automated ECG Beat Classification System Using Convolutional Neural Networks” 2016.