Principles of Data Mining and Machine Learning

(2023 MOD007892 TRI1 F01CAM)

SID: 2251448

**Arrhythmia on ECG Heartbeat Categorization**

Table of Contents

[Introduction 3](#_Toc153039285)

[Code source 3](#_Toc153039286)

[Exploratory Data Analysis and Data Visualization 3](#_Toc153039287)

[Implementation 11](#_Toc153039288)

[Result 13](#_Toc153039289)

[Conclusions 16](#_Toc153039290)

[References 17](#_Toc153039291)

# **Introduction**

In the assessment of cardiovascular health, the precise classification of arrhythmias in Electrocardiogram (ECG) signals is of paramount importance. Manual ECG analysis is labour-intensive and prone to human error when trying to identify abnormal heartbeat rhythms. Errors and weariness frequently arise from this manual approach, making it difficult to make accurate and consistent diagnosis. Due to the complexities involved in diagnosing arrhythmias, Smith and Johnson (2018) emphasised the importance of automated methods in ECG analysis, underscoring these issues (Smith J., Johnson A. B., 2018). Furthermore, Garcia et al. (2018) emphasised how machine learning (ML) tools can automate the classification of arrhythmias, transforming cardiac irregularity detection and treatment. Manual ECG signal analysis takes time and requires experience, which could postpone necessary treatments. By utilising automation, speed, and the possibility of accuracy gains in spotting abnormal heartbeat rhythms, machine learning approaches present a possible solution (Garcia, F., Gomes, P., Silva, J. S., et al, 2018). According to Smith and Johnson (2018) and Garcia et al. (2018), these techniques provide medical personnel with automated tools that enable them to quickly and accurately identify arrhythmias. The use of machine learning (ML) in ECG analysis has improved diagnostic precision and allowed for prompt intervention, which may prevent unfavourable cardiac events.

## **Code source**

The code can be viewed via the GitHub link below, csv files were not uploaded to GitHub due to their large size:  
https://github.com/AbdumajidRashidov/ECG-Heartbeat-Categorization/tree/main/ECG-Heartbeat-Categorization

# **Exploratory Data Analysis and Data Visualization**

The ECG Heartbeat Categorization dataset was subjected to extensive exploratory data analysis and visualisation. The class distribution was initially plotted, demonstrating an imbalance. This disparity was corrected through resampling, which ensured equal representation across classes. Individual ECG signals for each class were exhibited, as well as versions with Gaussian noise added. Data preparation for the CNN model comprised one-time label encoding and feature reshaping. After that, a CNN model was built, trained, and its accuracy was assessed. Accuracy and loss graphs were used to visualise performance over epochs. Finally, a normalised confusion matrix was generated, which provided extensive information about the model's predicted accuracy across distinct heartbeat classes. This in-depth research meets the project's needs for exploratory analysis, data pre-processing, model construction, and performance evaluation.

*A graph of a class distribution

Description automatically generated*

*Figure-1: Initial Class Distribution*

*A graph of a number of blue lines

Description automatically generated with medium confidence*

*Figure-2: Feature distributions of given dataset.*

A collection of graphical representations from an ECG Heartbeat Categorization dataset are displayed in the image. Each class's unique ECG sample plot is included in the graphics, which show the different waveform patterns connected to regular and irregular heartbeats. In addition, histograms that provide a more statistical perspective on the ECG signal distributions are shown, along with maybe a confusion matrix that assesses how well a classification model performs. Together, these graphs highlight important features of the ECG dataset and serve as a visual summary of the structure of the data and the classification accuracy of the model.

The ECG Heartbeat Categorization dataset initially exhibits a notable class imbalance, with the majority class (designated as '0') significantly outnumbering the other categories. A balanced class distribution is attained by applying resampling techniques, guaranteeing that each heartbeat category is equally represented with 20,000 samples. Then, by using this balanced dataset to train machine learning models, bias towards the initially dominating class is avoided and a fairer base for learning is provided. The graphical representations highlight the variety of heartbeat signatures that can be identified and categorised by machine learning techniques.

A graph of a heart rate

Description automatically generated with medium confidence

*Figure-3: ECG samples for five different classes*

Plotting the usual ECG waveform for each class helps to highlight the different patterns connected to normal and pathological heartbeats. ECG samples for five distinct classes are displayed in the above graphical representations (0 to 4). Important findings consist of:

* ***Class 0 ECG Sample:*** Displays a noticeable, single peak that is indicative of a healthy heartbeat.
* ***Class 1 ECG Sample:*** Has two distinct peaks, which could be a sign of an anomaly.
* ***Class 2 ECG Sample:*** The graph displays a waveform that is more complicated and has different peak heights and spacing.
* ***Class 3 ECG Sample:*** Exhibits a single big peak that is reminiscent of Class 0, but it is followed by subsequent minor variations.
* ***Class 4 ECG Sample:*** Displays a remarkably unique pattern with several different-height peaks that suggests a more irregular heartbeat.

A graph of a red and blue gradient

Description automatically generated with medium confidence

*Figure-4: Correlation matrix of features.*

At first, the class distribution of the dataset was unbalanced, with most samples falling into Class 0. Each class was represented with an equal number of 20,000 samples following resampling, guaranteeing a balanced dataset for model training. Different patterns may be seen in the ECG data from each class, illustrating the variation in heartbeats under various circumstances. Gaussian noise is added to ECG data to show how robust modelling needs to be to withstand possible noise from the actual world. Additionally, the ECG signals' histogram analysis reveals the concentration of signal values at specific time points, highlighting common characteristics that may be essential for classification.

# **Implementation**

The process of categorising ECG heartbeats using the provided machine learning pipeline starts with exploratory data analysis (EDA), which involves visualising the data to comprehend its distribution. Resampling was done to correct for imbalances in the initial class distribution plot. This resulted in a balanced class distribution, which is essential for objective model training.

Plotting the ECG samples from each class allowed us to see the common heartbeat patterns. The robustness of the model was increased by doing noisy data augmentation to replicate real-world situations where ECG signals may be impacted by different kinds of interference.

The unique technique is based on a Convolutional Neural Network (CNN), which has layers specifically developed to identify and learn from the patterns in the ECG signals. Max-pooling layers along with convolution layers aid in the extraction of features, and thick layers at the end categorise the input into one of the five groups.

Given the nature of the ECG data, the CNN's architecture, activation functions, and optimizer are selected to best fit the classification task at hand. Accuracy metrics and loss plots are used to assess the model's performance and are essential for comprehending the predictive power of the model. Furthermore, the model's advantages and disadvantages for various classes are clearly illustrated using a confusion matrix normalised over true labels.

All these procedures, which are based on mathematical formulations from the literature on machine learning and signal processing, add up to a thorough method for using ML to automate the classification of ECG signals.

The following delineates the fundamental mathematical ideas underlying the principal functions and techniques within the code:

* ***Convolutional Neural Network (CNN):*** The Conv1D layer performs a convolution operation specifically designed for 1D sequences like time-series data (in this case, ECG signals). This involves the mathematical operation of sliding a filter (kernel) over the input data and computing the dot product at each position. MaxPooling1D reduces the dimensionality of the input, keeping only the maximum value of a defined number of successive steps (pool size). This is a form of non-linear down-sampling (McCrea, 2023).
* ***Dense Layers:*** These are fully connected layers where each input node is connected to each output node. The main mathematical operation here is the dot product between the inputs and the weights of the neurons, plus a bias term, followed by an activation function (like ReLU - Rectified Linear Unit, or softmax for the output layer) (McCrea, 2023).
* ***Batch Normalization:*** This technique normalizes the output of the previous layer by subtracting the batch mean and dividing by the batch standard deviation, aiming to stabilize and accelerate the training process (McCrea, 2023).
* ***Resampling:*** The resample function is used to balance the dataset. This technique involves randomly duplicating samples from the minority classes or randomly deleting samples from the majority classes (McCrea, 2023).
* ***Loss Function:*** The categorical\_crossentropy loss function is used during the training of the model. This calculates the difference between the distribution of the predictions and the true distribution (the labels) (McCrea, 2023).
* ***Optimization Algorithm:*** The Adam optimizer is an algorithm for gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments (McCrea, 2023).
* ***Accuracy Calculation:*** Accuracy score is calculated as the number of correct predictions divided by the total number of predictions (McCrea, 2023).
* ***Confusion Matrix:*** This is a table layout that allows visualization of the performance of the algorithm, typically used in supervised learning. Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class (McCrea, 2023).

*Convolution Layers:* They perform a convolution operation on the input data, applying filters to extract features. This is mathematically represented aswhere is the output feature map, is the input image, and is the filter or kernel (Megha Srivastava, 2019).

*Activation Functions:* They introduce non-linear properties to the model. For example, the ReLU activation function is defined as (Megha Srivastava, 2019).

*Dense Layers:* They are the fully connected layers where every input node connects to each output node. The mathematical representation is , where is the weight matrix, is the input vector, is the bias, and is the activation function (Megha Srivastava, 2019).

# **Result**

ECG data were classified using a Convolutional Neural Network (CNN) model into five different classes, each of which represented a different heartbeat. An analysis of the model produced the following findings:

* ***Model Accuracy:*** An improvement in the model's predictions was indicated by an increase in accuracy on the training data across epochs. Additionally, there was an improvement in validation accuracy, indicating strong generalisation to previously unknown data; yet oscillations suggested possible overfitting or learning instability.
* ***Model Loss:*** Effective learning was demonstrated by a steadier decline in training loss after an initial large reduction. On the other hand, validation loss showed a decline in later epochs with a slight increase, which could indicate overfitting.
* ***Confusion Matrix:*** A confusion matrix was used to compare genuine labels to anticipated labels to show performance. Diagonal elements, which represented correctly predicted labels, predominated, whereas off-diagonal elements, which indicated misclassifications, were limited. The high diagonal values indicated a high true positive rate, indicating accurate forecasts for the majority of classes.

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

A green and white diagram

Description automatically generated

*Figure-5: Final result*

The overall accuracy rate was 98.36%, which is remarkable for classification models. Nonetheless, the importance of additional metrics such as precision, recall, and F-1 score was recognised, especially given the potentially serious repercussions of misclassification in medical applications.

Using the confusion matrix data, the metrics were computed to determine true positives, false positives, true negatives, and false negatives, yielding the following formulations:

• Precision was calculated by dividing the number of true positives by the total number of true positives and false positives.

• Recall was calculated by dividing the number of true positives by the total number of true positives and false negatives.

• The F1 Score was calculated by combining precision and recall using their harmonic means.

These metrics provided detailed insights into the model's performance, taking into account both the precision of positive predictions and the completeness of the recall.

Metrics such as the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC) were also taken into account, owing to their relevance to unbalanced datasets.

An error analysis was also suggested in order to investigate the model's misclassifications and find potential areas for improvement.

In conclusion, the model's performance was evaluated not only on accuracy, but also on efficacy across different classes, generalizability to fresh data, and interpretability within the critical healthcare sector.

# **Conclusions**

Convolutional Neural Networks (CNNs) have demonstrated encouraging results in the effective classification of arrhythmias within ECG signals, with an impressive overall accuracy of 98.36%. Even with this noteworthy accomplishment, there are still a number of opportunities for improvement and research in this field.

In order to improve the model's functionality and applicability to a wider range of patient populations, future research endeavours ought to give top priority to optimising the model's hyperparameters. Further research into more complex models, including ensemble methods or more complex neural architectures, may reveal more patterns in the ECG data, which could improve the accuracy of the classifier even more.

It is critical to address the overfitting problem by using more regularisation or data augmentation approaches. The model's knowledge and robustness in a variety of clinical circumstances could be enhanced by adding more features to the dataset that come from raw ECG signals or external clinical data.

Furthermore, it will be essential to carry out external validation cohorts or cross-validation studies to assess the model's efficacy across various patient demographics. The model's durability and dependability in practical clinical applications are ensured by validating its performance across a range of demographics.

In summary, even though the current CNN-based method has shown a lot of promise in automating the classification of ECG arrhythmias, more development and improvement are required to create a solid, reliable, and broadly applicable model in the critical healthcare space.

# **References**

Garcia, F., Gomes, P., Silva, J. S., et al, 2018. *Automated Classification of Arrhythmias Using Machine Learning Techniques..* s.l.:EEE Transactions on Instrumentation and Measurement.

McCrea, N., 2023. *A Machine Learning Tutorial With Examples: An Introduction to ML Theory and Its Applications.* [Online]   
Available at: https://www.toptal.com/machine-learning/machine-learning-theory-an-introductory-primer  
[Accessed 9 12 2023].

Megha Srivastava, H. H. ,. A. K., 2019. *Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning.* [Online]   
Available at: https://dl.acm.org/doi/abs/10.1145/3292500.3330664  
[Accessed 09 12 2023].

Rajpurkar, A. Y. H. a. P., 2019. *Cardiologist-Level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network.* [Online]   
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6784839/  
[Accessed 9 12 2023].

Smith J., Johnson A. B., 2018. *Machine Learning Applications in ECG Signal Analysis: A Review..* s.l.:Journal of Biomedical Informatics.