Diagnosing ‘silent’ heart attacks using ECG waveforms

Kristian Alikaj

Abstract

This project focuses on the early detection of silent heart attacks, one of the most pressing health concerns, through the innovative application of machine learning models to analyze electrocardiogram (ECG) waveforms. Silent heart attacks are a type of myocardial infarction that often go undiagnosed due to their asymptomatic nature, posing a significant health risk due to the delay in intervention and treatment which, consequently, leads to a higher risk of heart failure. The comprehensive Nightingale Open Science dataset, a meticulously curated dataset that aligns ECG waveforms with cardiac ultrasound data, served as the primary data source for this project. The primary objective of the study was the identification of Regional Wall Motion Abnormalities (RWMA), which are key indicators of prior myocardial infarctions. Various machine learning methodologies, including the more traditional Logistic Regression and Support Vector Machine (SVM), as well as the state-of-the-art Recurrent Neural Network (RNN), were evaluated in this research. The results demonstrated the immense potential of machine learning models in the detection of silent heart attacks through the analysis of ECG waveforms. Among the models evaluated, the RNN model stood out as it outperformed the other models in detecting RWMA due to its innate ability to capture sequential and time-dependent data. However, it is important to note that challenges such as data imbalance and overfitting were encountered during this study, highlighting the need for further refinement and optimization of the model.

Introduction

Heart attacks, also known as myocardial infarctions, are a major health concern worldwide. Every year, millions of heart attacks occur globally. However, a significant portion of these heart attacks, often referred to as "silent" heart attacks, remain undiagnosed. This is primarily because these heart attacks do not exhibit the typical symptoms associated with a heart attack, such as chest pain or discomfort, shortness of breath, and lightheadedness. As a result, patients often do not seek medical help, and thus, do not receive the necessary life-saving medications and treatments. This, in turn, can lead to devastating consequences, including death. Therefore, the early and accurate detection of prior myocardial infarction is of crucial importance as it allows for the provision of timely intervention and reduces the burden of cardiovascular disease on the healthcare system.

Electrocardiograms (ECGs) are a widely available and cost-effective diagnostic tool used in various healthcare settings for the detection of heart diseases. Clinicians have learned to identify some limited signs of prior heart attack on ECGs, such as Q-waves. However, these traditional methods still miss a substantial number of silent heart attacks. With the advent of machine learning techniques, there is now the potential to systematically mine ECG waveform data for more nuanced signals that may indicate prior heart attacks, thereby improving the detection rate of silent heart attacks.

In this study, the objective was to develop robust machine learning models, specifically Recurrent Neural Networks (RNNs), that are capable of detecting regional wall motion abnormalities (RWMA) from ECG waveforms, thereby serving as an indicator of silent heart attacks. RWMA, which signifies impaired movement of specific heart wall segments, is often a consequence of a previous myocardial infarction. By leveraging the Nightingale Open Science dataset, which links high-quality cardiac ultrasound data to ECG waveforms, algorithms were trained with the goal of potentially identifying a large fraction of undiagnosed prior heart attacks.

The accurate and scalable identification of silent heart attacks through ECG analysis could have a significant impact on patient care. It could enable wider access to life-saving medications and interventions, thereby reducing the risk of further complications and improving patient outcomes. Furthermore, the dataset used in this project, sourced from a public county health system, holds the promise of expanding the reach of such diagnostic tools to traditionally underserved patient populations, thereby promoting health equity.

Data and Methods

In this study, a dataset from the Nightingale Open Science platform was used. This dataset consists of a large number of ECG waveforms, specifically 48,788, collected from a diverse group of 13,438 patients (Pramanik et al., 2021). Each waveform represents a single 12-lead ECG and was collected by the Contra Costa Health Services (CCHS) county health system from 2013 to 2020 using the Philips TraceMasterVue ECG Management System. This ensures that the data is not only large in volume but also diverse, which is crucial for the development of a robust and generalizable machine learning model.

For the purposes of this research analysis, a specific subset of the data, known as the v0 dataset, was chosen to be the focus. This subset contains 5,000 ECGs from 5,000 unique patients, randomly selected from the larger v1 dataset. A unique feature of these patients was that they all had undergone a cardiac ultrasound within one year prior to their ECG, providing a unique opportunity to link ECG waveform data with cardiac ultrasound data. The ECG waveforms were saved as NumPy arrays, with each lead containing 5,500 sample points recorded at a sample rate of 500 Hz.

Also included in the dataset is a label indicating whether a regional wall motion abnormality (RWMA) was found in the patient's cardiac ultrasound report. In the v0 dataset, 9.6% of the ECGs were linked with a positive RWMA label (Pramanik et al., 2021).

The data is structured in a hierarchical file tree, with separate folders for CSV data and NumPy arrays. The CSV data includes details about the ECGs and RWMA outcomes, while the NumPy arrays contain the ECG waveforms for each lead.

The ECG waveforms were originally shared with us as XML files, which were parsed and stored in separate NumPy arrays for each lead. These arrays, which contain the waveforms for all the ECGs in the dataset, have a shape of ({number of ECGs},1,5500). Each waveform should ideally have 5,500 sample points, but if there are fewer, the array is filled with zeroes. The waveforms have a sample rate

of 500Hz and are measured in microvolts (uV).

The CSV file `lead-samples-count.csv` provides detailed information about each ECG, including the unique identifier, the NumPy index, and the number of samples for each lead.

Another crucial file, `rwma-outcomes.csv`, contains information about each of the 5,000 patients in the v0 dataset, specifically relating to their cardiac ultrasound results. Each patient had undergone a cardiac ultrasound within a year prior to their ECG, and this file records whether a regional wall motion abnormality (RWMA) was identified in the patient's cardiac ultrasound report. This data is crucial, as it serves as our ground truth for subsequent machine learning model development.

The data in this CSV file includes a unique identifier for each ECG (`ecg_id`), a flag indicating the presence of RWMA (`rwma`), and the NumPy index of the ECG (`npy_index`).

The ECG waveforms were originally received as XML files, which were parsed and stored in NumPy arrays. These arrays store the waveforms of each lead, and they include the waveforms for that lead for all the ECGs in the dataset. The shape of each array is ({number of ECGs},1,5500). It's expected that each waveform will have 5,500 sample points, and if it has fewer, the array is filled with zeroes.

The sample rate of these waveforms is 500 Hz and the amplitude units of the leads are in microvolts [uV].

In the CSV file `lead-samples-count.csv`, detailed information for each ECG is provided. This includes a unique identifier for each ECG (`ecg_id`), the NumPy index of the ECG (`npy_index`), and the number of samples for each lead, such as `MDC_ECG_LEAD_I`.

This data forms the foundation of the study, enabling us to delve deeper into the detection of 'silent' heart attacks using ECG waveforms (Pramanik et al., 2021).

Exploration of machine learning models for RWMA detection began with a simple Logistic Regression approach. The Logistic Regression model achieved an overall accuracy of 53.33%, with a precision of 0.92, recall of 0.53, and F1-score of 0.67 for the negative class (no RWMA), and a precision of 0.12, recall of 0.58, and F1-score of 0.20 for the positive RWMA class. This relatively poor performance suggested that a more sophisticated model would
be necessary to capture the complex patterns within the ECG waveform data.

Implementation

This research involved implementing and testing several machine learning algorithms using Jupyter notebooks. These notebooks detail each step of the process, from data preprocessing and feature scaling to model training, evaluation, and results visualization. The Jupyter notebooks are accessible in the repository: https://github.com/kris96tian/machine_learning_ecg/tree/main/Jupyter-notebooks-models

The study began with the importation of the necessary libraries. These included numerical computation (numpy), data manipulation and analysis (pandas), machine learning (sklearn), deep learning (torch), and data visualization (matplotlib and seaborn) libraries.

Next, the dataset was imported. It consisted of electrocardiogram (ECG) waveforms saved as NumPy arrays. Each of these arrays included waveforms for all the ECGs in the dataset, each waveform ideally having 5,500 sample points. Then, the ECG waveforms were processed, and feature scaling was applied. Feature scaling is a crucial step in machine learning as it standardizes the range of independent variables or features of data. This standardization makes the training process more efficient, helping the model to learn the weights of the features more effectively.

A function named `preprocess_data` was created to preprocess the data. This function was designed to load the data from the directory, and then stack the arrays into one large array. After loading and stacking the data, it was reshaped and permuted to fit the desired format. This preprocessing step ensured that the data was in the right format and ready for the model to process.

The next step was feature scaling, where a function named `feature_scaling` was used. This function scales the input arrays using the Normalizer from the sklearn.preprocessing library. The Normalizer works by scaling individual samples to have a unit norm.

Model Training

Once the data was preprocessed and scaled, it was split into training and test datasets. This was done using the `train_test_split` function from the sklearn.model_selection library. The dataset was split such that 80% of the data was used for training and the remaining 20% for testing.

The next step was to create the Recurrent Neural Network (RNN) model. The chosen RNN architecture consisted of a simple layer of LSTM (Long Short-Term Memory) cells, followed by a fully connected (dense) layer. The LSTM layer serves to process the input ECG waveforms, capturing their temporal dependencies. The dense layer then uses the output from the LSTM layer to make the final prediction of whether a RWMA is present or not.

The model was compiled with the Adam optimizer and binary cross-entropy loss function, suitable for the binary classification task at hand. The choice of metrics for model evaluation was accuracy, precision, recall, and F1-score, all of which are critical for assessing the performance of a binary classification model.

The model was trained using the training data over several epochs. An epoch is one complete pass through the entire training dataset. The number of epochs was chosen based on the model's performance on the validation dataset, which is a subset of the training dataset not used during the weight update process.

To prevent overfitting, early stopping was implemented. This technique stops the training process if the model's performance on the validation dataset does not improve for a certain number of consecutive epochs.

This number, known as the patience, was set based on experimental results.

Model Evaluation

After the model was trained, it was evaluated on the test dataset. The metrics obtained from this evaluation give an unbiased estimate of the model's performance on unseen data.

In addition to accuracy, precision, recall, and F1-score, the Receiver Operating Characteristic (ROC) curve was also plotted. The ROC curve is a graphical representation of the true positive rate against the false positive rate for the model. The area under the ROC curve (AUC-ROC) is a single-value metric that provides an overall measure of the model's discriminative ability.

Results and Discussion

The RNN model showed promising results in detecting RWMA from ECG waveforms. It achieved higher accuracy, precision, recall, and F1-score than the Logistic Regression model. Moreover, the AUC-ROC was also significantly higher, indicating better overall performance.

Despite these promising results, the model's performance on the positive class (RWMA present) was not as good as on the negative class. This may be due to the imbalance in the dataset, as only 9.6% of the ECGs were linked with a positive RWMA label. To address this issue, techniques such as oversampling the minority class or undersampling the majority class could be explored.

Another challenge faced was overfitting, as indicated by the divergence of the training and validation loss during the training process. This suggests that the model may be too complex and is learning the noise in the training data. To mitigate overfitting, regularization techniques such as dropout or l1/l2 regularization could be implemented.

In conclusion, this study demonstrates the potential of using machine learning, specifically RNNs, to detect silent heart attacks through ECG waveforms. With further refinement and validation on larger and more diverse datasets, this approach could become a valuable tool in the early detection and management of cardiovascular disease. This could potentially transform the way we diagnose and treat heart disease, ultimately leading to improved patient outcomes and a reduction in the global burden of cardiovascular disease.

References

Pramanik, Rajiv, Bhumil Shah, Anna Roth, Honga Wei, Ted Castillo, Katie Lin, Sachin Shah, et al. 2021. “Diagnosing ’Silent’ Heart Attack Using ECG Waveforms: A Nightingale Open Science Dataset.” Nightingale Open Science. https://doi.org/10.48815/N54W2V.