Predicting Cardiac Wellness: Using a Multi-Layer Perceptron on ECG Data

Benzon Carlitos Salazar

CS732: Project Report

Introduction

The exploration of classification algorithms on ECG data to detect heart disease lies in the ability of classification algorithms to analyze and interpret complex patterns within the ECG signals. Several compelling motivations drive this pursuit: firstly, classification algorithms enable the early identification of heart abnormalities and arrhythmias by discerning patterns indicative of various heart conditions. This capability is crucial for timely intervention and effective management of cardiovascular health. Additionally, the manual analysis of ECG signals is not only time-consuming but also demands specialized expertise, two important things I unfortunately do not have. Classification algorithms automate this process, providing an efficient and scalable approach to analyzing extensive volumes of ECG data, offering a practical solution for the challenges posed by manual ECG reading.

Furthermore, these algorithms offer an objective and consistent analysis of ECG data, mitigating the potential for human error or subjectivity in results interpretation. This consistency is vital for reliable and reproducible diagnostics. Lastly, different classification algorithms can be tailored to identify specific types of arrhythmias or abnormalities, enhancing the diagnostic capacity of the entire system and providing detailed insights into the nature of the heart condition.

Crash Course Anatomy on ECG Signals

An Electrocardiogram (ECG) is a graph representing the voltage versus time of the heart's electrical activity, measured using electrodes typically placed on the chest. ECGs are crucial for diagnosing various heart diseases, including arrhythmias, which contribute significantly to global mortality, as per the World Health Organization.

The PQRST waves in a standard ECG represent distinct electrical events during one cardiac cycle. The P wave represents atrial depolarization, the QRS complex represents ventricular depolarization, and the T wave represents ventricular repolarization. The ST segment between the QRS complex and the T wave represents the time between ventricular depolarization and repolarization, crucial for identifying myocardial infarction or other cardiac issues.

Prior Works

In 2022, Darmawahyuni et al. proposed a one-dimensional Convolutional Neural Network (1D-CNN) for ECG rhythm and beat classification. Notably, their approach simplified the classification process by employing a single deep learning architecture for both rhythm and beat features, achieving high accuracy across nine rhythm and fifteen beat classes. The low computational requirements of the 1D-CNN model enhance its suitability for real-time and cost-effective applications in ECG devices [1].

In 2023, Ramkumar et al. presented an ensemble classifier for arrhythmia detection, emphasizing the significance of their AD-Ensemble SVM-NB-RF method. The study aimed to create a universal model applicable to the general population, employing ensemble classifiers (SVMs, Naive Bayes, and random forest) for arrhythmia detection. The two-stage arrhythmia classification utilized Residual Exemplars of Local Binary Pattern (RELBP) for pre-processing ECG signals [2].

In 2019, Alfaras et al. introduced a fast machine learning model for ECG-based heartbeat classification and arrhythmia detection. Their proposed ensemble of Echo State Networks (ESNs) addressed the challenge of long computation times for classifiers, showcasing advantages in speed and memory over traditional methods. Recurrent connections and parallel computing architecture were key components of their methodology [3].

Dataset Overview

The dataset chosen for this study is the MIT-BIH Arrhythmia Database, which includes 48 half-hour excerpts of two-channel ambulatory ECG recordings. These recordings were obtained from 47 subjects studied by the BIH Arrhythmia Laboratory between 1975 and 1979.

The dataset's unique characteristics include:

- **Comprehensive Annotation:** Two or more cardiologists independently annotated each record, resolving disagreements to obtain computer-readable reference annotations for each beat, totaling approximately 110,000 annotations. This detailed annotation ensures the reliability and richness of the dataset.
- **Random Selection for Diversity:** Twenty-three recordings were randomly selected from a set of 4000 24-hour ambulatory ECG recordings collected from a mixed population of inpatients (about 60%) and outpatients (about 40%) at Boston's Beth Israel Hospital. The remaining 25 recordings were chosen to include less common but clinically significant arrhythmias that would not be well-represented in a small random sample. This intentional diversity enhances the dataset's representativeness.
- **Digitization and Resolution:** The recordings were digitized at 360 samples per second per channel with 11-bit resolution over a 10 mV range. This high-resolution digitization allows for detailed analysis of the ECG signals.

- **Long-Term Availability:** The MIT-BIH Arrhythmia Database has been freely available since September 1999, and additional signal files were posted in February 2005. This long-term availability ensures accessibility for researchers and practitioners over an extended period.

This dataset provides a robust foundation for training and testing classification models, offering a diverse and well-annotated set of ECG recordings representative of various cardiac conditions.

Choice of Algorithm: Multilayer Perceptron

Reason for Multilayer Perceptron

The selection of the Multilayer Perceptron (MLP) as the algorithm for ECG data analysis stems from the intricate nature of ECG signals. ECG signals often exhibit highly complex and non-linear patterns, and MLPs, being neural networks, possess the capacity to learn and model these intricate mappings between input features (representing ECG signal characteristics) and output classes (indicating various arrhythmia types). This inherent flexibility makes MLPs particularly suitable for capturing the nuanced relationships within high-dimensional and intricate ECG data that might pose challenges for linear classifiers.

MLPs are adept at automatically learning hierarchical representations of features from raw ECG data during the training process. Unlike manual feature design, where relevant features are identified and crafted by experts, MLPs autonomously identify and learn these features, a particularly beneficial characteristic when dealing with the diverse and complex nature of ECG signals. Additionally, MLPs are known for their flexibility, enabling them to adapt to different types of data. Given the variability in ECG signals across individuals and populations, MLPs can adjust their internal parameters during training to accommodate these variations, enhancing the algorithm's adaptability and generalization.

Furthermore, the choice of scikit-learn's MLP implementation is driven by its capability to handle datasets of varying sizes. Whether dealing with a small dataset for initial experiments or a larger dataset for more robust model training, MLPs can scale effectively to accommodate diverse data sizes.

What is a Multilayer Perceptron?

The Multilayer Perceptron (MLP) is a supervised learning algorithm designed to approximate a function mapping input features to output targets using a labeled dataset. Unlike linear models such as logistic regression, MLPs can model non-linear relationships in the data by incorporating one or more non-linear layers, known as hidden layers, between the input and output layers. This distinctive ability to capture non-linear relationships is crucial when dealing with complex datasets like ECG signals.

Multilayer Perceptron Optimization

MLP in scikit-learn can be trained using different optimization algorithms, each with its unique characteristics. The optimization algorithms include Stochastic Gradient Descent (SGD), Adam, and L-BFGS.

- **Stochastic Gradient Descent (SGD):** This algorithm updates parameters using the gradient of the loss function with respect to the parameters. The learning rate (α) controls the step-size in the parameter space search. While effective, its performance can be sensitive to the choice of the learning rate.
- **Adam Optimizer:** Similar to SGD, Adam is a stochastic optimizer but incorporates adaptive learning rates. It automatically adjusts the update step based on adaptive estimates of lower-order moments, providing robustness against variations in the learning rate. This study opted for Adam during the training process.
- **L-BFGS Solver:** L-BFGS is a solver that approximates the Hessian matrix and its inverse for parameter updates. It is particularly suitable for small to medium-sized datasets. While computationally intensive, it can converge faster on certain problems.

The choice of the optimization algorithm depends on the specific characteristics of the dataset and the desired trade-off between computational efficiency and model performance. In this study, Adam was chosen for its adaptive learning rates and demonstrated effectiveness in training MLPs for ECG data classification.

Methodology

Data Processing

1. **Download MIT-BIH Arrhythmia Database from PhysioNet:** The database was downloaded to obtain 48 half-hour excerpts of two-channel ambulatory ECG recordings.
2. **Convert WFDB files to CSV:**
    a. Split at R-peaks: The recordings were split at R-peaks to isolate individual cardiac cycles.
    b. Read QRS Complex: The QRS complex, a critical component of the ECG signal, was read for analysis.
    c. Resample at 125 Hz: The data was resampled to a standard frequency of 125 Hz for consistency.
    d. Normalize mV Readings: Voltage readings were normalized to a 0-1 range for uniformity.
    e. Discard Erroneous Values: Any erroneous values were discarded for data integrity.
    f. Reduce Classifications to Normal/Abnormal: The dataset was simplified to binary classifications of Normal or Abnormal.

3. **Generate Training, Validation, and Test CSV files via random shuffling:**
    a. Training: 60% of records were allocated to the training set.
    b. Validation: 20% of records were allocated to the validation set.
    c. Testing: 20% of records were allocated to the testing set.
4. **Train, Validate, and Test the model**

Feature Extraction

Instead of using annotations to find beats, R-peak detection was employed. Heartbeat classifications from the annotations were reduced to Normal and Abnormal, appended to each heartbeat record.

Results

Validation Phase

During the validation phase, the performance metrics of the trained Multilayer Perceptron (MLP) model were evaluated, showcasing impressive outcomes:

- **Test Accuracy:** 98.09%
- **Precision:** 98.87%
- **Recall:** 98.60%
- **Specificity:** 95.71%
- **F1 Score:** 98.73%

The exceptionally high accuracy, precision, recall, specificity, and F1 score collectively suggest that the model performs exceptionally well across various evaluation criteria. The Receiver Operating Characteristic (ROC) curve further emphasizes its effectiveness, with a high Area Under the Curve (AUC) of 0.97. This indicates the model's robust ability to distinguish between positive and negative instances, validating its efficacy.

Testing Phase

Similar outstanding results were observed during the testing phase, affirming the model's generalization to unseen data:

- **Test Accuracy:** 98.18%
- **Precision:** 98.85%
- **Recall:** 98.73%
- **Specificity:** 96.11%
- **F1 Score:** 98.79%

The consistency in performance metrics between the validation and testing phases is a positive indicator. The model demonstrates not only high accuracy but also remarkable precision, recall,

specificity, and F1 score during both phases. This consistency signifies the model's reliability and its ability to generalize effectively to new and unseen data.

These exceptional results underscore the proficiency of the final MLP model in accurately predicting arrhythmias from ECG data. The high level of precision and recall, coupled with the model's consistent performance across validation and testing phases, reinforces confidence in its diagnostic capabilities and its potential for practical utility in the realm of ECG analysis and cardiovascular health.

Conclusion

Challenges Faced

The implementation of the ECG data classification project encountered several challenges throughout its development:

Firstly, the pre-processing phase posed a significant hurdle due to the unfamiliarity with processing waveform signals. While having extensive experience with unstructured text and tabular data, dealing with the intricacies of waveform signals proved time-consuming.

Secondly, the inherently noisy nature of ECG data complicated the accurate identification of each PQRST wave. To address this challenge, a strategic decision was made to switch to the MIT-BIH Arrhythmia dataset, renowned for its cleanliness and comprehensiveness. This change aimed to ensure the reliability and accuracy of the dataset, laying a robust foundation for subsequent analyses.

A hardware limitation further impacted the project. The initial intention was to implement TensorFlow’s DNNClassifier, a Deep Neural Network Classifier. However, due to the absence of the necessary hardware and GPUs for running the classifier on large amounts of data, an alternative solution was sought. Consequently, the Multilayer Perceptron (MLP) Classifier was adopted, ensuring computational feasibility within the existing hardware constraints.

Potential Future Work

Several avenues for potential future work and enhancements to the project includes;

Firstly, incorporating both TensorFlow and PyTorch into the project can open doors to advanced deep learning techniques and models. The unique strengths and features of each framework offer opportunities for exploring more sophisticated algorithms, potentially leading to improved model performance.

Expanding the dataset from half-hour ECG readings to a more prolonged duration, such as an hour, emerges as a promising avenue. This extension provides an opportunity to capture additional temporal patterns and variations in cardiac activity, contributing to a more

comprehensive understanding of heart rhythms and potentially enhancing the model's ability to generalize.

Considering the contemporary context, updating the dataset to include records from 2020 onwards, especially those involving patients affected by COVID-19, introduces a relevant and current perspective. Analyzing ECG data from individuals with COVID-19 may unveil unique cardiac signatures associated with the disease, thereby contributing to a deeper understanding of its impact on cardiovascular health.

The implementation of ensemble learning methods, such as combining predictions from multiple models, stands out as a potential enhancement. Ensemble techniques often lead to improved model performance by mitigating individual model biases and capturing a more comprehensive representation of underlying patterns. Introducing such methods could enhance the robustness and generalization of the model, especially in handling diverse and complex ECG data.

Overall Conclusion

The achieved exceptional metrics, including a remarkable accuracy of 98%, demonstrate the model's proficiency in accurately predicting arrhythmias from ECG data. With these metrics, there is a high level of confidence in the model's ability to precisely identify and classify arrhythmias. The success of the final model signifies the practical utility in providing reliable predictions, contributing to the field of ECG analysis, and facilitating early detection and intervention for cardiovascular health.

Citation

[1] Darmawahyuni, Annisa et al. “Deep learning-based electrocardiogram rhythm and beat features for heart abnormality classification.” *PeerJ. Computer science* vol. 8 e825. 25 Jan. 2022, doi:10.7717/peerj-cs.825

[2] Ramkumar, M., et al. "Ensemble classifier fostered detection of arrhythmia using ECG data." *Medical & Biological Engineering & Computing* (2023): 1-14.

[3] Alfaras, Miquel, Miguel C. Soriano, and Silvia Ortín. "A fast machine learning model for ECG-based heartbeat classification and arrhythmia detection." *Frontiers in Physics* (2019): 103.