Report 1 - Sami Ghannam

I. Introduction

Electrocardiogram (ECG) signals are commonly used to analyze the electrical activity of the heart. Automatic ECG heartbeat classification is an important task in biomedical signal processing, as it can help detect cardiac arrhythmias efficiently and consistently.

The objective of this practical work is to perform ECG heartbeat categorization using a supervised machine learning approach. A public dataset from Kaggle based on the MIT-BIH Arrhythmia Database is used. A baseline classification model is implemented and its performance is compared with results reported in the original literature.

II. Dataset Description

The dataset used in this project is the MIT-BIH Arrhythmia dataset available on Kaggle. It contains ECG heartbeat segments extracted from long-term ECG recordings and pre-processed into fixed-length signals.

Each sample corresponds to a single heartbeat represented by 187 numerical values describing the ECG waveform. The last column represents the class label. The dataset includes five heartbeat classes labeled from 0 to 4.

The data is provided as two separate files:

- `mitbih_train.csv` for training
- `mitbih_test.csv` for testing

The dataset is highly imbalanced, with the majority of samples belonging to class 0 (normal heartbeats), while other classes representing arrhythmias contain significantly fewer samples.

III. Model Implementation

A multi-class Logistic Regression model was implemented as a baseline classifier. This model was chosen for its simplicity, fast training time, and ease of interpretation.

Before training, all input features were standardized using feature scaling in order to ensure that each feature contributes equally to the learning process. The model was trained on the training set and evaluated on the independent test set.

The performance of the classifier was evaluated using accuracy, precision, recall, and F1-score metrics.

IV. Results

The Logistic Regression model achieved a test accuracy of 91.47%. The results show very good performance for the majority class, while lower recall values are observed for minority classes.

This behavior can be explained by the strong class imbalance present in the dataset and the linear nature of the classifier, which makes it less effective at capturing complex patterns associated with rare arrhythmias.

V. Hyperparameter Experimentation

To study the effect of hyperparameters, several values of the regularization parameter were tested for the Logistic Regression model. The results showed that moderate regularization values provide the best balance between underfitting and overfitting.

This experiment highlights the importance of hyperparameter tuning, even for simple baseline models.

VI. Comparison with the Original Paper

Previous studies on the MIT-BIH Arrhythmia dataset report higher classification performance, often above 97% accuracy. These results are typically achieved using more advanced models such as convolutional or recurrent neural networks, along with specific techniques to address class imbalance.

In comparison, the model used in this work is a simple baseline classifier without explicit imbalance handling, which explains the lower performance observed.

VII. Conclusion

In this practical work, an ECG heartbeat classification system was implemented using a Logistic Regression model. Despite its simplicity, the model achieved reasonable performance and provided insight into the challenges of ECG classification.

This work demonstrates that baseline machine learning models are useful reference points and that more advanced approaches are required to achieve state-of-the-art results on imbalanced biomedical datasets.