Beat Classifier for PPG Signals

Drmic A., Javadi M. and Shala M.

I. INTRODUCTION

Photoplethysmography (PPG) signal classification is a crucial task in biomedical engineering, particularly for diagnosing and monitoring heart conditions. This project aims to develop a beat classifier for PPG signals to classify each beat into Normal (N), Supraventricular (S), and Ventricular (V). Accurate classification of these beats is vital for detecting arrhythmias and other cardiac anomalies, which can significantly impact patient outcomes.

In this assignment, we use a dataset comprising PPG signals from 105 patients with varying sampling frequencies. The primary goal is to preprocess these signals, train a classifier, and evaluate its performance in distinguishing between the different types of heartbeats.

The subsequent sections of this report will detail the materials and methods used, including data preprocessing techniques, model architecture, and training procedures. We will then present the results obtained from our experiments, followed by a discussion on the implications of these findings. Finally, we will conclude with a summary of our work and suggestions for future research directions.

II. MATERIALS AND METHODS

A. *Dataset*

The dataset used in this project Consists of PPG recordings from 105 patients, with two different sampling frequencies: more specifically, 62 patients were recorded with f_S = 128Hz and the remaining 43 patients with f_S = 250Hz collected to facilitate the classification of heartbeats into three distinct categories: Normal (N), Supraventricular (S), and Ventricular (V).

But the dataset is predominantly composed of Normal (N) beats, with only a small fraction being anomalous. This creates a significant class imbalance as illustrated in Figure [1] that was addressed using class weighting and erasing all the samples with all normal beats.

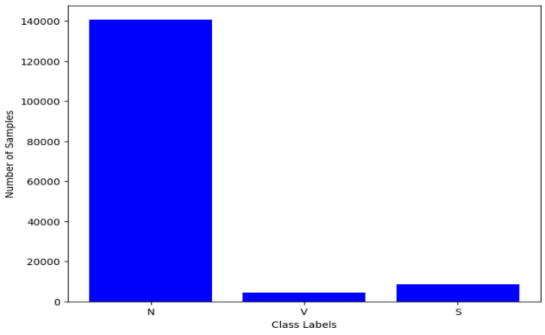

Figure 1. Class Distribution

B. *Data Splitting and Stratified Sampling*

In this project, we divided our dataset of heartbeats from 105 patients into **training**, **validation**, and **testing sets** using a **stratified sampling** approach to maintain class distribution. Patients were grouped into those with predominantly 'N' annotations and those with fewer 'N' annotations. Each group was then independently split into training, validation, and testing sets while maintaining the 70%, 20%, and 10% proportions, respectively. This approach ensured that each subset was representative of the overall dataset.

The training data was used to train the model, with **batching** and **shuffling** to improve learning efficiency. The validation data was used during training to monitor the model's performance and apply **early stopping** if necessary. Finally, the testing data was used to evaluate the model's performance and assess its generalization capability. This method ensured that the data subsets were balanced and representative, leading to a robust and well-generalized model.

C. *Signal Resampling*

To ensure uniformity across the dataset, signals originally recorded at 128 Hz were **resampled to 250 Hz**, the higher of the two sampling frequencies. This standardization is essential for consistent signal processing and analysis.

D. *Normalization*

To prepare the data for classification, we first normalized both the training and validation sets. Normalization is beneficial as it scales the data to a consistent range, improving the convergence speed of learning algorithms and ensuring that each feature contributes equally to the analysis.

E. *Noise Cancellation*

A critical step in preprocessing was to address noise, which can significantly impair our task. Initially, we applied a **Butterworth bandpass filter** with cutoff frequencies at 0.25 Hz and 10 Hz to mitigate noise and smooth the signal. Although this method effectively reduced general noise, it failed to eliminate certain high amplitude oscillations observed upon manual inspection of the training set samples. Consequently, these oscillations necessitated the exclusion of affected samples from the training dataset.

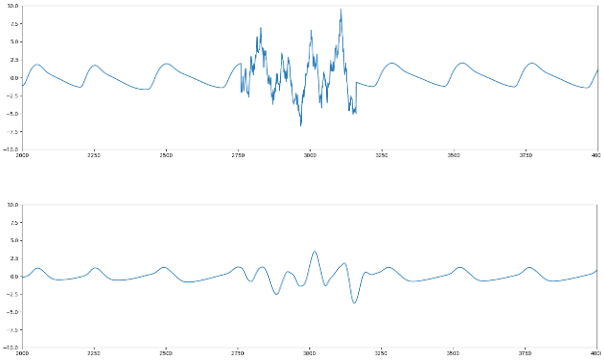

Figure 2. Signal Before and After Filtering

F. Peak segmentation

Subsequently, we segmented each peak by creating windows centered around each peak, extending 250 samples on either side. This approach allows us to capture the temporal dependencies surrounding each peak, which is hypothesized to contain beneficial information for classification. Peaks within the first and last 40 samples of each signal were excluded to avoid boundary effects.

G. Valid Segments

To further address the issue of high amplitude oscillations, we calculated the peak-to-peak value (Max - Min) for each segment. Segments exhibiting excessively high peak-to-peak amplitudes were deemed as containing outliers and were consequently discarded from the training dataset and the valid segments were achieved. This multi-step preprocessing ensured a cleaner and more reliable dataset for subsequent classification tasks.

H. Metrics

Accurate classification is crucial, particularly in cases involving abnormalities, where misclassifying an abnormal peak (S or V) as normal can have significant consequences. Given the imbalanced nature of our dataset, with many normal labels, accuracy alone is not an adequate metric. Therefore, we must employ more appropriate metrics to better achieve our goal of minimizing false negatives and improving overall classification performance.
For imbalanced classification problems like ours, where one class significantly outweighs the others, metrics beyond accuracy are necessary to provide a comprehensive evaluation of model performance. Here are several metrics that we explored in this project:

1. Precision

Precision measures the accuracy of positive predictions made by the model. It is calculated as the ratio of true positive predictions to the total number of positive predictions (both true positives and false positives). A high precision indicates that when the model predicts a positive result, it is likely to be correct. In the context of our project on PPG signal classification, precision helps us understand how reliably the model identifies abnormal peaks (S or V) without incorrectly labeling normal peaks.

$$Precision = \frac{TP}{TP + FP}$$

2. Recall

Recall measures the ability of the model to correctly identify all actual positive instances. It is calculated as the ratio of true positive predictions to the total number of actual positives (true positives and false negatives). High recall indicates that the model is good at capturing all positive instances, which in our case, means correctly identifying abnormal peaks (S or V) in PPG signals.

$$Recall = \frac{TP}{TP + FN}$$

3. Specificity

Specificity measures the ability of the model to correctly identify all actual negative instances. It is calculated as the ratio of true negative predictions to the total number of actual negatives (true negatives and false positives). In our context, specificity helps us understand how well the model avoids incorrectly classifying normal peaks as abnormal, thus ensuring that only true negatives (normal peaks) are identified as such.

$$Specificity = \frac{TN}{TN + FP}$$

4. Adaptive Metric Adjustment Strategy

In addition to evaluating standard metrics, we employed a dynamic adjustment strategy based on empirical penalties and thresholds. This approach allowed us to fine-tune our model's performance, particularly crucial for handling imbalanced datasets like our classification of PPG signals. By adjusting metrics such as True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) using a penalty factor and specific thresholds, we aimed to optimize classification accuracy while minimizing misclassifications. This iterative approach ensured that our model continuously improved its ability to accurately identify abnormal peaks (S or V) amidst varying data

distributions, enhancing its reliability in practical medical applications.

5. *Weighted Accuracy*

Weighted accuracy is a crucial metric in multi-class classification problems, particularly when dealing with imbalanced datasets. In this context, our model classifies beats into three classes with significantly varying class distributions (136193 samples for N, 6364 for S, and 5696 for V). To account for this imbalance, weighted accuracy assigns different levels of importance to each class, based on their inverse frequencies. This approach ensures that the less frequent classes (S and V) are given proportionally higher weights, thus preventing the model from being biased towards the majority class (N). By calculating weighted accuracy from the confusion matrix, we obtain a more balanced measure of the model's performance. This provides a comprehensive evaluation, highlighting the model's effectiveness across the entire dataset.

$$Weighted\ Accuracy = \frac{\sum TP}{\sum (TP + FN)}$$

CLASSIFICATION

In this study, the task of classifying PPG signals was approached using various deep learning models tailored to capture distinctive aspects of the data. Given the automatic feature extraction capabilities inherent in deep learning methods, our focus was on leveraging segmented peaks derived from the data preprocessing stage as inputs to these models.

Deep Learning Models

To comprehensively evaluate the classification of PPG signals, we implemented several prominent model architectures, each selected for its ability to capture nuanced patterns within temporal data.

1. *VGG-Style CNN*

A modified VGG-style Convolutional Neural Network (CNN) was employed to effectively capture temporal dependencies and subtle features present in PPG signals. This architecture was specifically chosen to address the challenges posed by an imbalanced dataset. The model was optimized using the **Adam optimizer** with a conservative learning rate of $1e^{-6}$, ensuring stable and gradual convergence.

2. *LSTM Model*

The Long Short-Term Memory (LSTM) model, renowned for its proficiency in modeling temporal dependencies, was employed to discern intricate patterns within the PPG signals. Optimized with the Adam optimizer at a learning rate of $1e^{-3}$, this model was designed to facilitate accurate classification.

3. *Bidirectional LSTM Model*

Building upon the LSTM architecture, the Bidirectional LSTM model was implemented to enhance learning by processing data in both forward and backward temporal directions. Similarly optimized with the Adam optimizer at a learning rate of $1e^{-3}$, this model aimed to capitalize on bidirectional information flow for improved classification performance.

4. *1D CNN Model*

Tailored to extract localized features from PPG signals through convolutional operations, the 1D CNN model provided a robust framework for feature extraction and classification. Compiled using the Adam optimizer, this model was configured to effectively capture spatial patterns within the temporal domain of the signals.

III. RESULTS

In our project, we prioritized the evaluation of two key metrics: **Weighted Accuracy** and **Recall**. These metrics were chosen based on their critical importance to our task, which involves both multi-class and binary classification scenarios. Weighted Accuracy provides a comprehensive assessment of overall predictive performance, adjusting for the varying distribution of classes within our dataset. Meanwhile, Recall measures the model's capability to correctly identify all instances of a specific class, essential for accurate classification across different types of PPG signals. High values of Recall highlight the correctly identified abnormal peaks. Results on the test set are as follows.

A. Binary

Model/Metric	Weighted Accuracy	Recall
VGG-Style	69.12%	0.8695
LSTM	70.62%	0.9052
Bi-LSTM	68.43%	0.8902
1D CNN	67.80%	0.8475

B. Multiclass

Model/Metric	Weighted Accuracy	Recall
VGG-Style	92.32%	0.8350
LSTM	68.75%	0.8801
Bi-LSTM	65.47%	0.8650
1D CNN	64.12%	0.8250

C. Confusion Matrix

The confusion matrix presented summarizes the performance of a classification model designed to categorize instances into two classes: N and non-N or three classes: N, S, and V for classification of **VGG CNN**. Each cell in the matrix represents the count of instances where the predicted class aligns with the actual class.

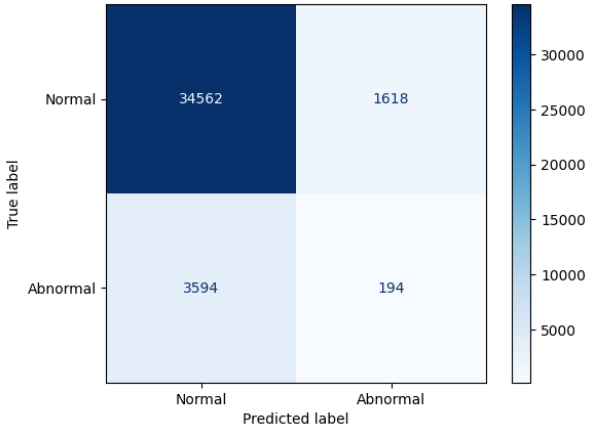

Figure 2. Confusion Matrix for Binary

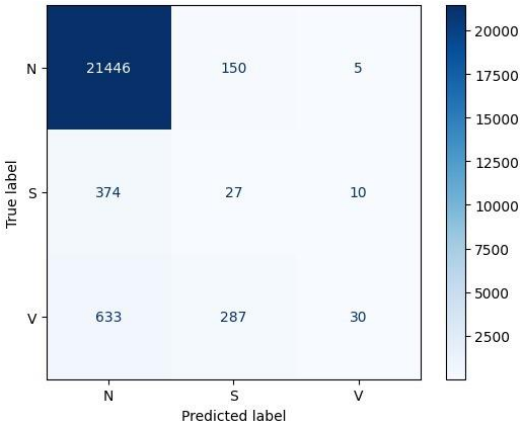

Figure 3. Confusion Matrix for Multiclass

D. Confidence Estimation

In our convolutional neural network model for classifying the three classes, the final layer employs a **softmax** activation function. This layer outputs a **probability distribution** over the three classes for each beat. The softmax probability corresponding to class V provides a direct estimate of the model's confidence in classifying a given beat as V. A higher softmax probability indicates greater confidence in the prediction. For instance, if the softmax output for the i-th beat shows a probability of 0.9 for class V, this suggests a high level of confidence that the beat is indeed a non-normal type 2 beat. Conversely, a lower probability, such as 0.4, would imply less confidence in this classification. By setting a threshold (e.g., 0.7), we can filter out predictions that fall below this level, thereby focusing on classifications where the model's confidence is robust. This thresholding mechanism helps in ensuring that only predictions with a high degree of certainty are considered, thereby improving the reliability of the model in practical applications.

In the binary classification of beats, the softmax layer is replaced by a **sigmoid** activation function in the output layer, which provides a single probability score indicating the confidence that a beat is non-normal. This score ranges from 0 to 1, where values close to 1 suggest high confidence that the beat is non-normal, and values close to 0 indicate high confidence that the beat is normal.

For example, if the sigmoid output for a beat is 0.85, this implies a strong confidence that the beat is non-normal. Conversely, a score of 0.15 would indicate a high confidence that the beat is normal. By setting an appropriate threshold, such as 0.5 or higher, we can determine the cutoff for classifying a beat as non-normal, ensuring that only predictions with sufficient confidence are considered.

IV. DISCUSSION AND CONCLUSION

The results from our study demonstrate the effectiveness of various deep learning models in classifying PPG signals into Normal (N), Supraventricular (S), and Ventricular (V) beats. In the binary classification task, all models—VGG-Style CNN, LSTM, Bi-LSTM, and 1D CNN—performed similarly, with LSTM achieving the highest weighted accuracy of 70.62% and recall of 0.9052. This indicates robust performance in distinguishing between normal and non-normal beats, crucial for preliminary detection of arrhythmias. However, when extended to multiclass classification, the VGG-Style CNN emerged as the top performer with a weighted accuracy of 92.32% and recall of 0.8350. This model showcased superior ability in classifying beats across all three categories. The LSTM and Bi-LSTM models, although strong in binary classification, demonstrated lower performance in the multiclass setting, suggesting potential challenges in effectively handling the increased complexity of distinguishing between three classes.

While our study yielded promising results in overall classification accuracy, we acknowledge that **further refinement is needed**, particularly in the classification of

minority classes such as Supraventricular (S) and Ventricular (V) beats. The performance metrics for these classes indicate room for improvement. Future efforts could focus on augmenting the dataset with more balanced representation across all classes or employing advanced techniques to mitigate class imbalance, thereby enhancing the model's ability to accurately classify all types of heartbeats.

In conclusion, our study underscores the importance of employing advanced deep learning architectures, particularly the VGG-Style CNN, for accurate classification of PPG signal beats in clinical contexts. The high weighted accuracy and recall achieved by the VGG-Style CNN in multiclass classification affirm it is the best among the trained models. Moving forward, further improvements can be explored by enhancing model interpretability, optimizing preprocessing techniques tailored to varying sampling frequencies. Additionally, expanding the dataset to include a more diverse patient population would provide a more robust evaluation of their generalizability and efficacy in clinical practice.