Technische Universität Kaiserslautern
Institute of
Cognitive Integrated Sensor Systems
Department of Electrical and Computer Engineering

Feature Extraction and Classification of ECG signal

Sharath Lokesh
Tanmay Kota
Winter Semester, 2022 - 2023

Semester Project in Sensor Signal Processing
Supervisor: Prof. Dr.-Ing. Andreas König

Overview

1. Introduction
    1.1. ECG – Electrocardiogram
    1.2. Cardiac Cycle

2. Methodology
    2.1. Dataset
    2.2. Data Pre-processing
    2.3. Feature Extraction and Selection
    2.4. Dimensionality reduction
    2.5. Model design and training
    2.6. Model testing and evaluation

3. Conclusion and Future Scope

4. References

5. Code

Introduction - ECG – Electrocardiogram

- Electrocardiogram (ECG) is a medical test that records the electrical activity of the heart from 12 different angles. [1]
- The 12 leads are obtained by placing electrodes on the chest and limbs of the patient and recording the electrical signals produced by the heart.
- Due to the 3D nature of the heart we need 12 leads. Each lead records the electrical activity of the heart from a specific angle.
- 12 Lead ECG consists of 6 limb leads and 6 chest leads, which are obtained from 10 electrodes (4 attached to limbs and 6 attached across the chest).
- Limb leads look at the heart in the vertical plane obtained by 3 electrodes RA, LA, and LL. The electrode on the right leg is an earth electrode (helps to provide a stable baseline for the ECG by providing a reference point that is not affected by the electrical activity of the heart).
- Chest Leads (precordial leads) view the chest in the horizontal plane, which uses 6 electrodes V1 to V6.

Introduction - ECG – Electrocardiogram

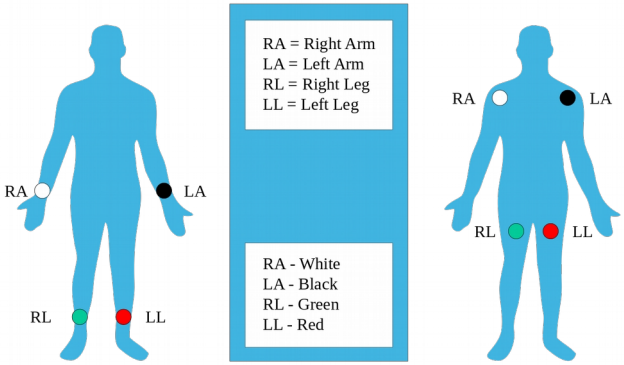


Fig. 1. Placement of Limb Electrodes [1]

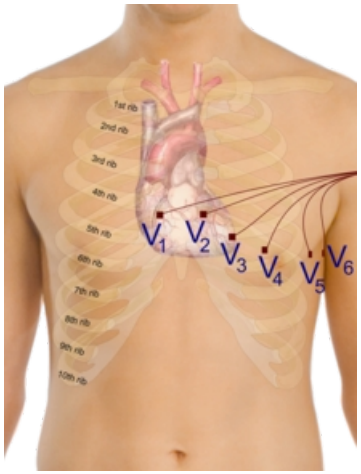


Fig. 2. Placement of Chest Electrodes [1]

Limb Electrode signals	Chest Electrode Signals
Lead I - right arm and left arm electrodes	V1 - electrical activity in the right ventricle
Lead II - right arm and left leg electrodes	V2 - electrical activity in the septum (the wall that separates the left and right ventricles)
Lead III - left arm and left leg electrodes	V3 - electrical activity in the anterior (front) wall of the left ventricle
aVR - right arm and a virtual electrode	V4 - electrical activity in the anterior-lateral (front-side) wall of the left ventricle
aVL - left arm and a virtual electrode	V5 - electrical activity in the lateral (side) wall of the left ventricle
aVF - left leg and a virtual electrode	V6 - electrical activity in the lateral (side) wall of the left ventricle, farther away from the heart than V5

Table 1. ECG signals from 12 leads

Introduction - Cardiac Cycle

The cardiac cycle is the sequence of events that occur during one heartbeat, as shown by the electrical signals generated by the heart. The cardiac cycle consists of several phases, including:

- **P wave**: The first wave of the ECG signal, indicating the contraction of the atria, which forces blood into the ventricles.
- **PR segment**: A flat line between the P wave and the QRS complex, indicating the delay in the electrical signal as it passes through the atrioventricular node (AV node) before reaching the ventricles.
- **QRS complex**: A large wave indicating the contraction of the ventricles, which pumps blood out of the heart and into the arteries.
- **ST segment**: A flat line between the QRS complex and the T wave, indicating the plateau phase of the cardiac muscle contraction.
- **T wave**: A small wave indicating the relaxation of the ventricles as they prepare for the next heartbeat.

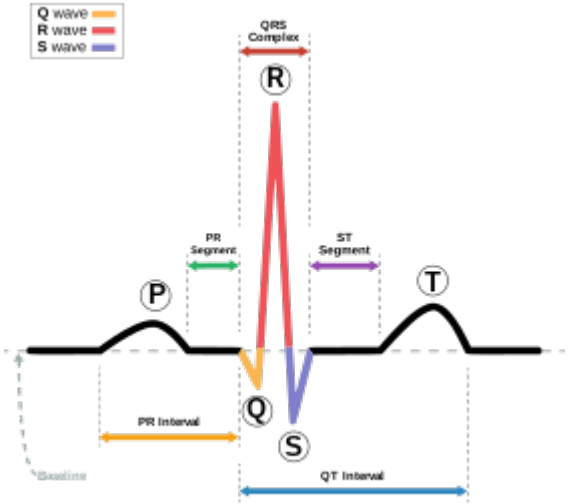


Fig. 3. Cardiac Cycle [1]

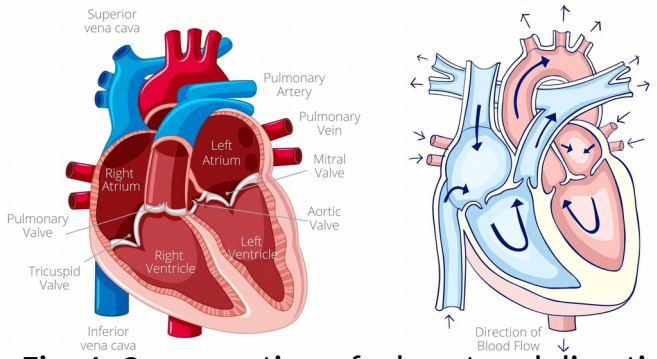


Fig. 4. Cross-section of a heart and direction of blood flow

Methodology

Problem statement

Investigate an open sources of ECG data and existing features and their implementation for this medical data processing. Design a SENSIG-system with the common steps of feature(s) computation, dimensionality reduction, and classification. Methods can be any from the presented spectrum of the lecture, but no deep learning solution.

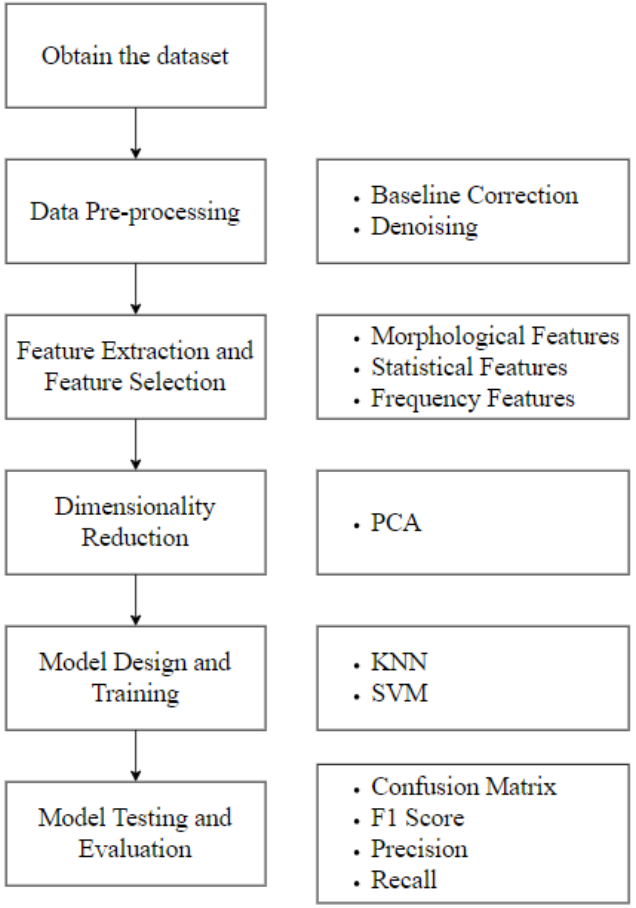


Fig. 5. Flow for ECG signal classification

Methodology - Dataset

- The PTB-XL ECG dataset is a large dataset of 21837 clinical 12-lead ECGs from 18869 patients of 10-second length. The PTB-XL dataset is preferred for arrhythmia research due to its large, diverse, and annotated 12-lead ECG recordings, making it a standard benchmark for machine learning in ECG analysis.
- The waveforms - 16-bit precision and a sampling frequency of 500Hz. For the user's convenience, it also has down-sampled versions of the waveform data at a sampling frequency of 100Hz (used in this project).
- The dataset consists of 5 classes:

NORM: Normal ECG (9528 #signals).

MI: Myocardial Infarction a.k.a. heart attack (5486 #signals).

STTC: ST/T Change, ST, and T wave changes may represent cardiac pathology or be a normal variant (5250 #signals).

CD: Conduction Disturbance. Your heart rhythm is the way your heart beats. Conduction is how electrical impulses travel through your heart, which causes it to beat. Some conduction disorders can cause arrhythmias, or irregular heartbeats (4097 #signals).

HYP: Hypertrophy, Hypertrophic cardiomyopathy (HCM) is a disease in which the heart muscle becomes abnormally thick (hypertrophied). (2655 #signals).

Methodology - Dataset

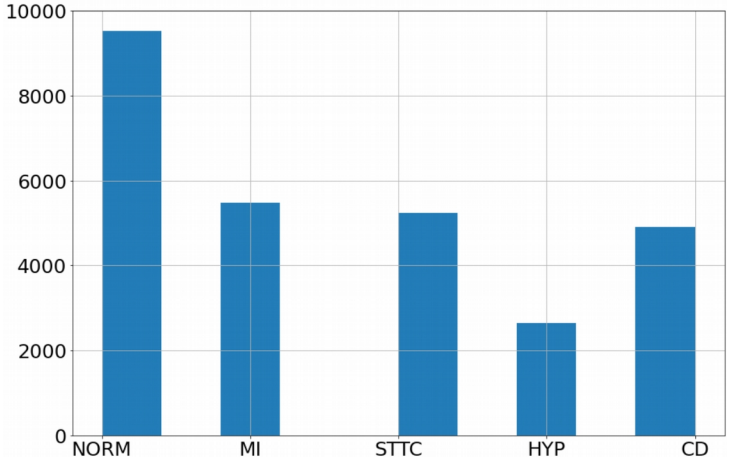


Fig. 6. Histogram for class distribution

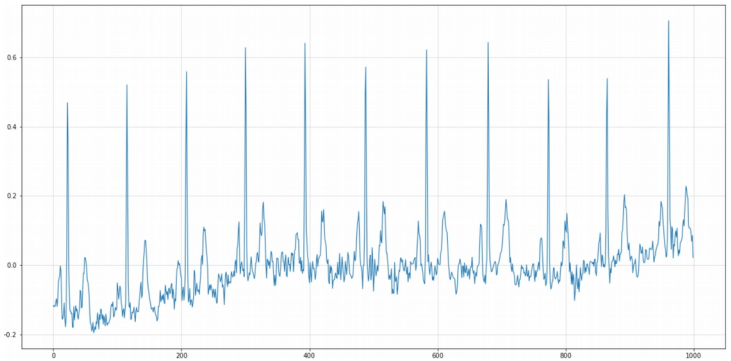


Fig. 7. Lead I Raw ECG signal sample

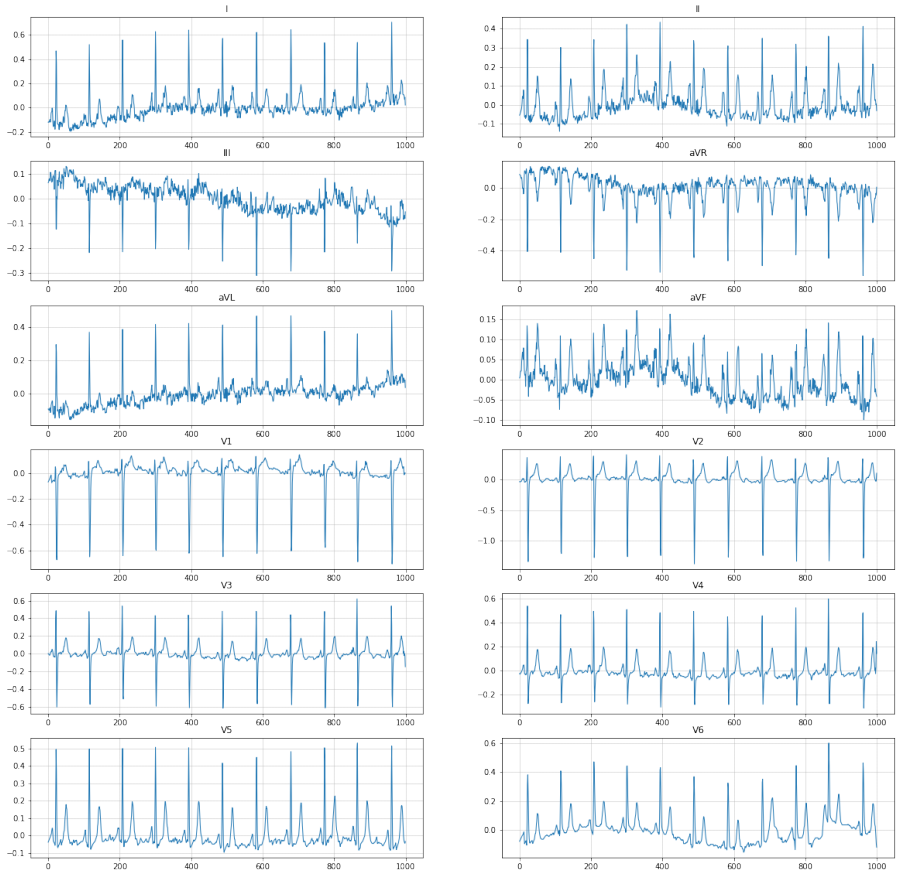


Fig. 8. 12 lead Raw ECG signal

Methodology – Data Pre-processing

- Electrocardiographic (ECG) signals are often contaminated with different types of noise and baseline drift.
- ECG signals are usually contaminated by the noise of diverse recourses and forms: motion artifact, 50/60Hz power line interference, and baseline drift caused by respiration [5][4].
- Techniques of wavelet packet, multi-wavelet, lifting wavelet, and bionic wavelet, were proposed and applied in ECG de-noising with the improvement of wavelet theory. In these techniques, a large number of experiments are required to take to select the appropriate scales and thresholds [5][4].
- FIR and IIR filters can effectively filter the high frequency but have difficulty in removing the additive noise which has the same frequency band as the ECG signal. [5][4].
- Zhongguo Liu, Jinliang Wang, and Boqiang Liu suggest a morphological filtering approach is simple, fast, and real-time in processing, and it keeps the ECG signal shape unchanged while removing the noise [4].
- Two basic morphological operators: erosion and dilation.
- Opening and closing are derived operators defined in terms of erosion and dilation.

Methodology – Data Pre-processing

Consider a pre-processed signal of length **N** and a structuring element of length **M**, such that **N>>M**.

- **Erosion**: Erosion is a morphological operation that shrinks and thins the boundaries of foreground objects in an image while removing small, isolated regions. In the context of ECG signals, erosion can be used to remove noise and baseline wander from the signal.

- **Dilation**: Dilation is the opposite of erosion and is used to expand and thicken the boundaries of foreground objects in an image. In ECG signals, dilation can be used to enhance and detect the QRS complex of the ECG signal.

Methodology – Data Pre-processing

- **Opening**: Opening is a morphological operation that consists of an erosion followed by a dilation. It can be used to remove small objects from the foreground while preserving the shape and size of larger objects. In ECG signals, the opening can be used to remove noise and baseline wander while preserving the larger features of the signal.

- **Closing**: Closing is the opposite of opening and consists of a dilation followed by an erosion. It can be used to fill in small gaps in the foreground while preserving the overall shape and size of the objects. In ECG signals, the closing can be used to fill in gaps in the QRS complex and enhance the detection of the complex.

- A combination of opening and closing functions can be used to remove baseline wandering and noise from the ECG signal [4]. Zhongguo Liu ， Jinliang Wang, and Boqiang Liu use an opening - closing – closing - opening (OC_CO) filter with a varying width of the structural element to retrieve a processed ECG signal from the raw signal.

Methodology – Data Pre-processing

- Fig. 9 represents the design of the OC_CO filter [4] which uses a linear structuring element (horizontal structuring element) of width K [6].
- Fig. 10 represents the Morphological filter for baseline wandering. The raw ECG signal f0 is the input for the first OC_CO filter with k = 0.11 Fs (where Fs is the sampling rate of the signal). The output f_b is the signal without the QRS complex. f_b is passed through an another OC_CO filter with k = 0.27 Fs. The resultant signal f_c lacks QRS complex, P, and T waves.
- This f_c signal is the baseline drift of the ECG which is subtracted from the original ECG signal to get the baseline corrected signal f_bc.
- Fig. 11 depicts the denoising of this baseline corrected signal by passing it through the OC_CO filter with k = 0.3 Fs.

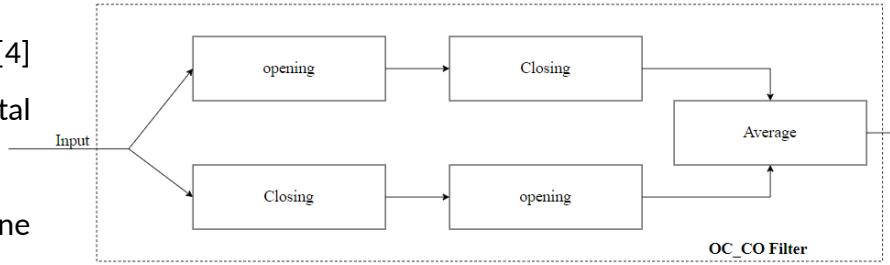


Fig. 9. OC_CO Filter

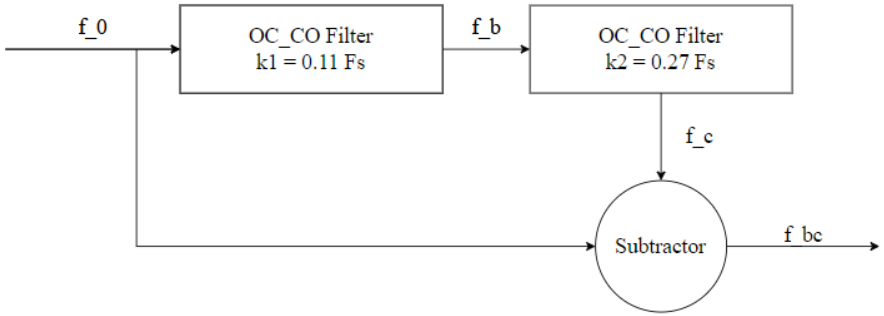


Fig. 10. Morphological Filter for baseline correction

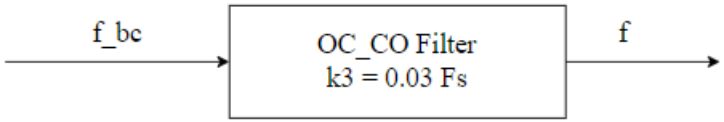
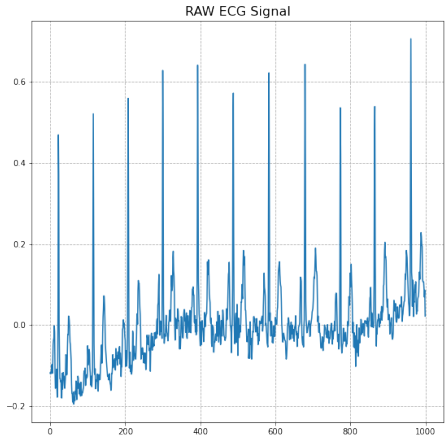
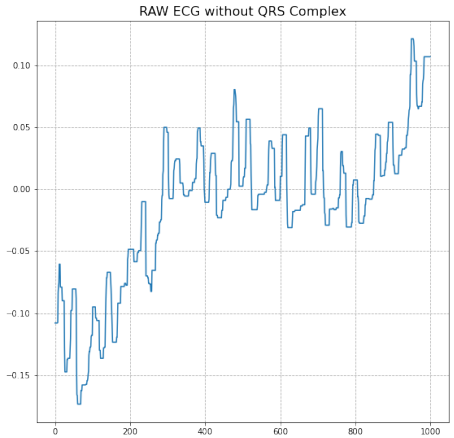
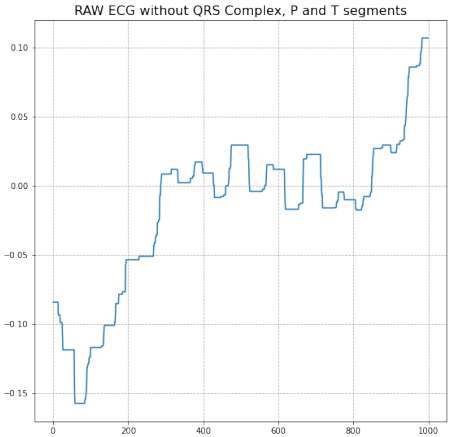


Fig. 11. Morphological Filter for ECG denoising

Methodology - Data Pre-processing







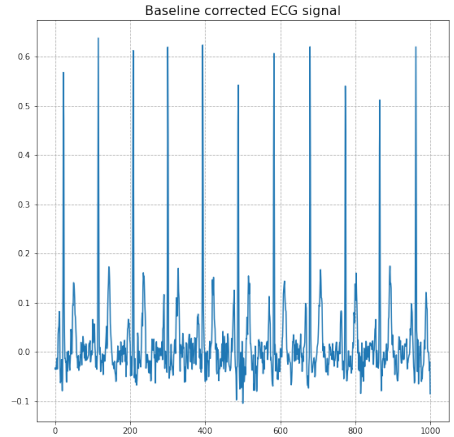
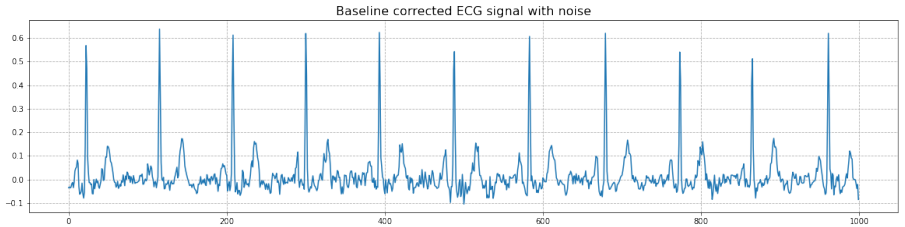


Fig. 12. Base line correction using OC_CO filters



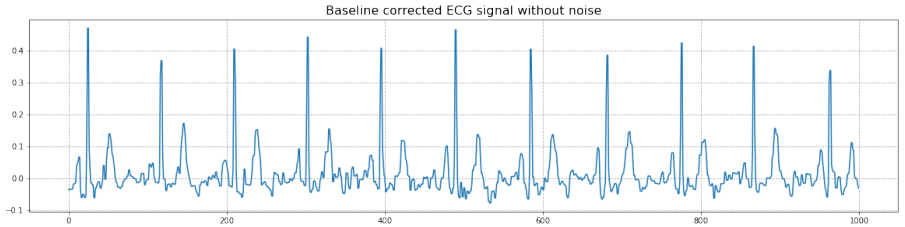


Fig. 13. Noise suppression using OC_CO filters

Methodology – Feature Extraction and Selection

- Features are classified into (i) Morphological features, (ii) statistical features, and (iii) frequency domain features.
- **Morphological Features**: QRS complex duration, P-wave duration and amplitude, T-wave duration and amplitude, S-amplitude, R-peak amplitude, RR interval duration, ST interval, Q amplitude (10 features). [8]
- The morphological features are extracted using the NeuroKit2 python library [9] [10]. This library uses peak detection along with gradient techniques or continuous wavelet transform or discrete wavelet transform for the detection of peaks and amplitudes in an ECG.

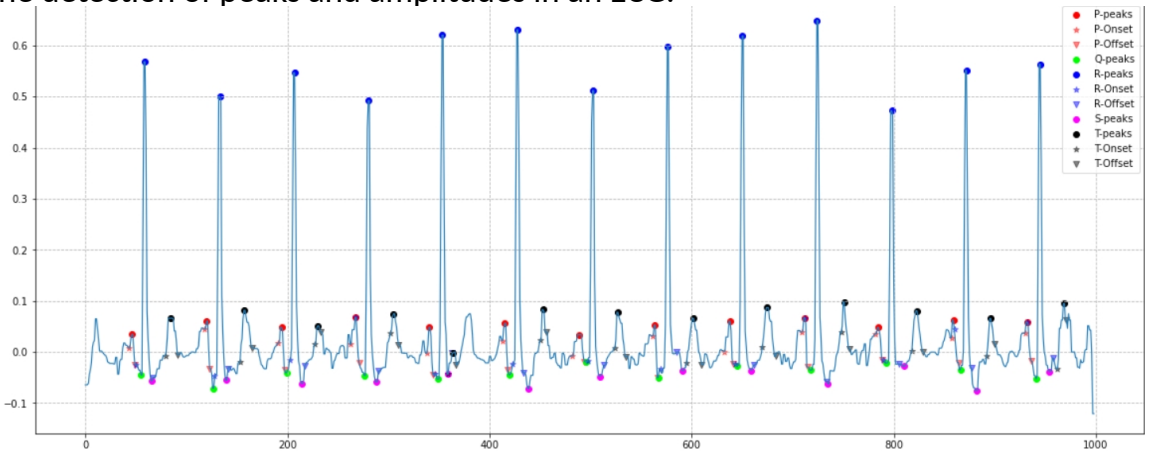


Fig. 14. ECG with morphological features

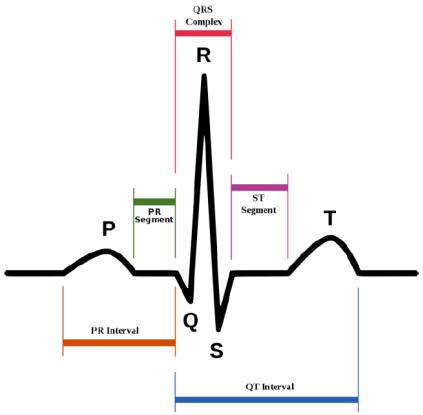


Fig. 15. ECG wave and Intervals

Methodology – Feature Extraction and Selection

Statistical Features: Variance, Mean, Skewness, Standard deviation, Spectral entropy, and Kurtosis. (6 features). [8]

- **Variance**: The variance is a measure of the spread of the ECG signal values around their mean. It can indicate the variability of the heart rate and rhythm.
- **Mean**: The mean of the ECG signal is simply the average of all the values. It can provide information about the baseline of the signal.
- **Skewness**: Skewness measures the asymmetry of the ECG signal. Skewness can provide information about the shape of the ECG waveform.
- **Standard deviation**: The standard deviation is a measure of the spread of the ECG signal values around their mean. It can provide information about the variability of the heart rate and rhythm.
- **Spectral entropy**: Spectral entropy is a measure of the complexity of the ECG signal. It is calculated from the spectral density of the signal and can provide information about the frequency content of the signal.
- **Kurtosis**: Kurtosis measures the "peakedness" of the ECG signal. A higher kurtosis value indicates a more peaked distribution of values, while a lower kurtosis value indicates a flatter distribution. Kurtosis can provide information about the shape of the ECG waveform.

Methodology – Feature Extraction and Selection

Frequency Domain Features: are extracted from the Power Spectral estimation using Welch's method [12] (11 features).

- In the Welch method, the ECG signal was split into overlapping segments. These segments were split into L data segments of length M, with overlap by D points. The window method is applied to overlapping segments and the periodogram is computed using Discrete Fourier Transform. Each segment is averaged to produce the estimate of the PSD. The segments are multiplied by a window function.
- The study by S. K. Shrikanth Rao, M. H. Kolekar, and R. J. Martis, have used Hamming window with a window width of 256 and several overlapping points as 128 and 512 FFT points.
- Frequency domain features extracted from PSD using Welch's method in ECG classification provide valuable information about the distribution of signal frequencies in the electrocardiogram. These features help capture the underlying rhythmic patterns and frequency components in the ECG signal, which can be essential for diagnosing cardiac conditions.

Parameter	Description
ULF Power	Absolute power of ULF
VLF Power	Absolute power of VLF
LF Peak	Peak frequency of LF
LF Power	Absolute power of LF
LF Power	Relative power of LF in nu
LF Power	Relative power of LF
HF Peak	Peak frequency of HF
HF Power	Absolute power of HF
HF Power	Relative power of HF in nu
HF Power	Relative power of HF
LF/HF	Ratio of LF to HF

Fig. 16. Frequency domain features

Methodology – Feature Extraction and Selection

- Frequency domain measurement calculates power in 4 frequency bands namely ULF (Ultra Low Frequency - up to 1 Hz), VLF (Very Low Frequency - 1 to 5 Hz), LF (Low Frequency - 5 to 20 Hz), and HF (20 to 50 Hz). [12]
- The frequency domain features shown in Fig. 16 are computed.

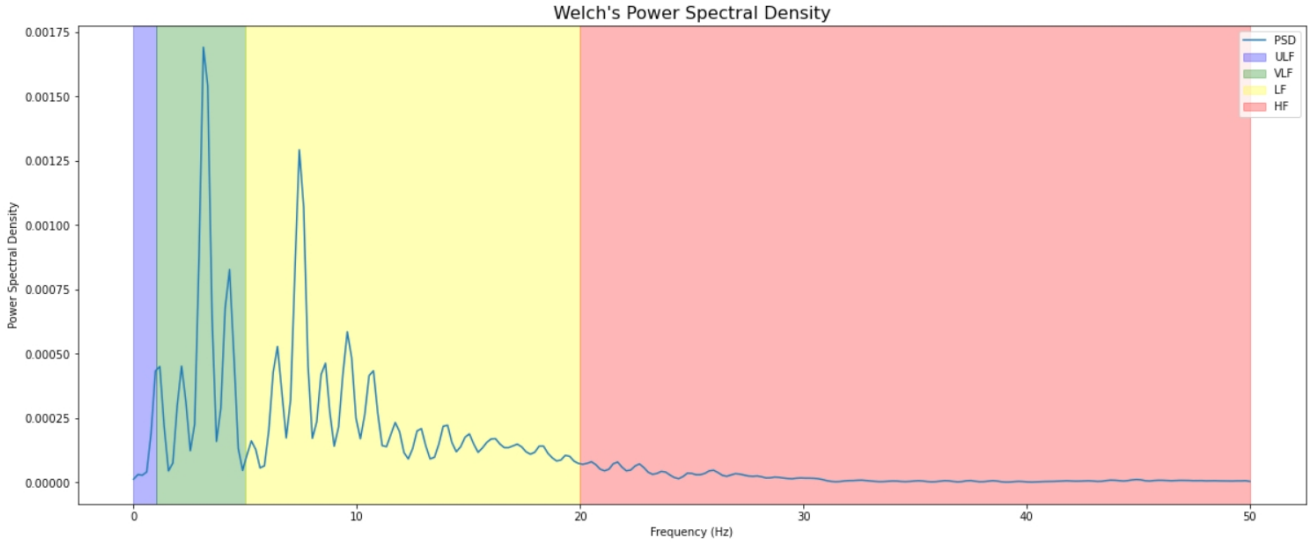


Fig. 17. PDS of an ECG signal using Welch's method

Methodology – Dimensionality reduction using PCA

- Concatenating all the features, the shape of the input is (21837, 27, 12). In some signals, we were not able to extract morphological features using the neurokit2 library. After discarding those signals the input matrix dimensions were (15602, 27, 12).
- Since the 12 signals in a single id correspond to one value in the label we concatenate the features from these 12 signals and the resulting dimensions were (15602, 324).
- The corresponding class distribution is STTC: 2231; MI: 1793; NORM: 1770; CD: 1721; HYP: 1600. (Fig. 17)
- From Fig. 17 it is clear that the dataset is unbalanced and can affect the model training and evaluation. So, we convert the dataset into a balanced dataset by considering approximately 1600 samples per class.
- The dimension of the balanced dataset is (6402, 324).

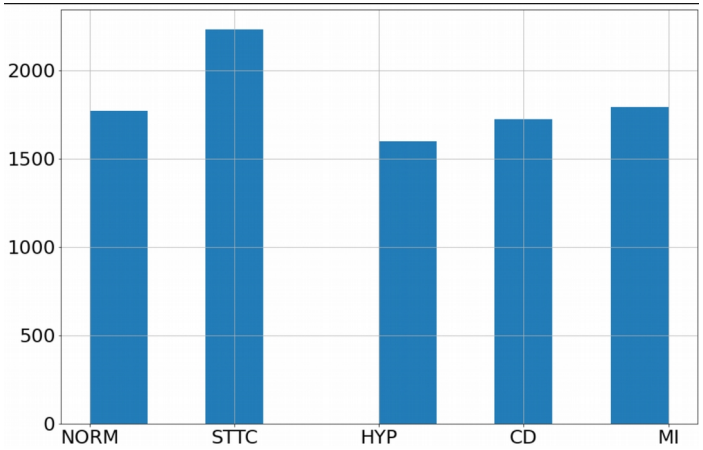


Fig. 18. Class distribution after Feature Cleaning

Methodology – Dimensionality reduction using PCA

- Each signal has 324 features to classify them into one or more classes.
- Not all the features in the datasets generated are important for training the machine learning algorithms. Some features might be irrelevant and some might not affect the outcome of the prediction. Ignoring or removing these irrelevant or less important features reduces the burden on machine learning algorithms. This is known as dimensionality reduction.
- The advantages of dimensionality reduction are: Increase in computational efficiency, Better visualization, Avoids overfitting, Data Compression, etc.
- The prominent dimensionality reduction techniques, Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are investigated on some popular Machine Learning (ML) algorithms, Decision Trees, and Support Vector Machine (SVM). [13]
- In this project, we use PCA to reduce the dimension of the feature space.
- PCA is a statistical procedure that uses an orthogonal transformation. PCA can be used for the examination of the relationships among a group of variables. Hence, it can be used for dimensionality reduction.

Methodology – Dimensionality reduction using PCA

PCA involves the following steps: [13]

- **Standardize the data**: The first step in PCA is to standardize the data, which means that each feature has zero mean and unit variance. This is important because PCA is sensitive to the scale of the features.
- **Calculate the covariance matrix**: The next step is to calculate the covariance matrix of the standardized data.
- **Calculate the eigenvectors and eigenvalues**: The eigenvectors and eigenvalues of the covariance matrix are calculated. The eigenvectors represent the principal components, while the eigenvalues represent the amount of variance explained by each principal component.
- **Choose the number of principal components**: The next step is to choose the number of principal components to keep. This can be done by looking at the eigenvalues and choosing the top k eigenvalues that explain most of the variance in the data.
- **Project the data onto the principal components**: Finally, the data is projected onto the selected principal components to obtain a lower-dimensional representation of the data.

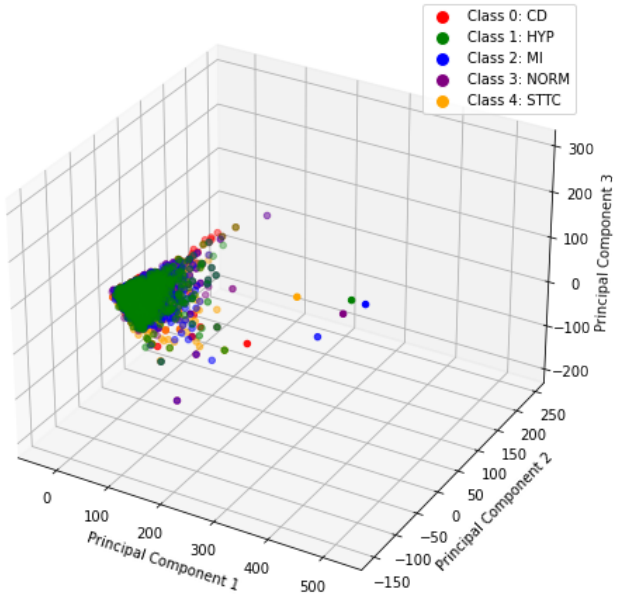


Fig. 19. Visualizing 3 principal component features

Methodology – Model design and training

- The data is split into training and testing with a split of 80:20%. That is 5150 samples are used for training and 1252 for testing.
- To ensure uniform distribution of the classes in train and test data a technique known as the stratified split is adopted.
- The annotation of the data set is a multilabel annotation, which means a signal belongs to at least one class. For example, if the signal can be classified as STTC and HYP the corresponding label will be ('STTC', 'HYP').
- The SVM and KNN classifiers would not be able to interpret the categorical labels as the input. Therefore, the classifier would not be able to distinguish between different classes and provide accurate predictions. One hot encoding is a technique used to convert categorical data into a format that can be used for machine learning algorithms.
- For example, if the output label is ('STTC', 'HYP') the one hot encoded format would be [01001] (['CD', 'HYP', 'MI', 'NORM', 'STTC']) where each bit represents the classification of the signal.

Methodology – Model design and training

Many Machine learning techniques that are used for ECG signal Classification. One of the popular techniques are SVM (Support Vector Machines), DT (Decision Trees), KNN (K-Nearest Neighbors), etc. This project focuses on exploring two such techniques namely, SVM and KNN.

- **Support Vector Machines** (SVM) is a widely used supervised learning algorithm for classification, regression, and outlier detection. SVM is a discriminative classifier that aims to find the hyperplane that maximizes the margin between the classes in the feature space [14][15].
- SVM tries to find the best-separating hyperplane in the feature space that can separate the positive and negative examples with the largest margin possible. The margin is defined as the distance between the hyperplane and the closest points of the positive and negative examples. The hyperplane that maximizes the margin is considered the optimal one, as it is expected to perform well on unseen data. [14]

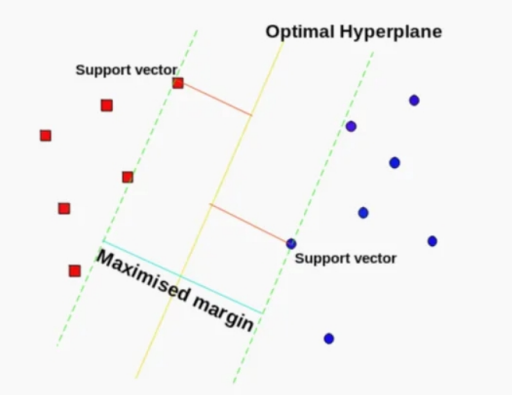


Fig. 20. Optimal Hyperplane using the SVM algorithm [14]

Methodology – Model design and training

- If the data cannot be linearly separated, SVM uses a technique called kernel trick, which maps the original feature space into a higher-dimensional space where the data becomes separable. The kernel function computes the dot product of the feature vectors in this higher-dimensional space without actually computing the coordinates of the data in this space. [15]
- In the case of multi-class classification, SVM can be used to solve the problem either in a one-vs-one or one-vs-rest approach.
- In the case of multilabel classification, SVM can be extended to handle multiple labels by using a binary relevance approach. In this approach, we treat each label independently as a binary classification problem and train a separate SVM for each label. During prediction, the probability of each label is predicted using the corresponding SVM, and the labels with the highest probabilities are assigned.

Methodology – Model design and training

- **K-Nearest Neighbors** (KNN) is a popular classification algorithm used for both binary and multiclass classification. It is a non-parametric algorithm that relies on the similarities between instances for making predictions.
- The basic idea behind KNN is to find the K nearest neighbors to a new data point in the training set based on a distance metric, and then to classify the new data point based on the class labels of those neighbors. The most common distance metric used in KNN is Euclidean distance, but other metrics such as Manhattan distance, Minkowski distance, and Hamming distance can also be used [16].
- The value of K in KNN is a hyperparameter that needs to be tuned. A small value of K (e.g., K=1) will lead to overfitting, as the algorithm may simply memorize the training set. On the other hand, a large value of K (e.g., K=10) may lead to underfitting, as the algorithm may not be able to capture the underlying structure of the data.

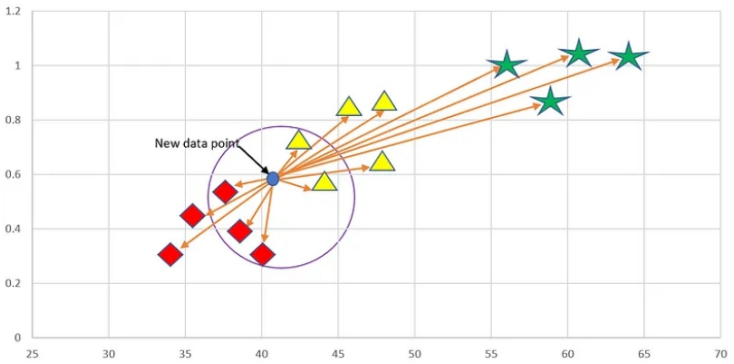


Fig. 21. KNN with K=5 [16]

Methodology – Model design and training

- While designing KNN the optimal value of the hyperparameter K has to be determined.
- The value of K can be obtained by evaluating the model for different values of K with respect to a metric (grid search algorithm) [17].
- The F1 score is used as the evaluation metric in this project.
- Fig. 21 shows the variation of the F1 score with respect to the K value, considering all 324 features.
- Fig. 22 shows that the model has the highest value of F1 score when K = 3

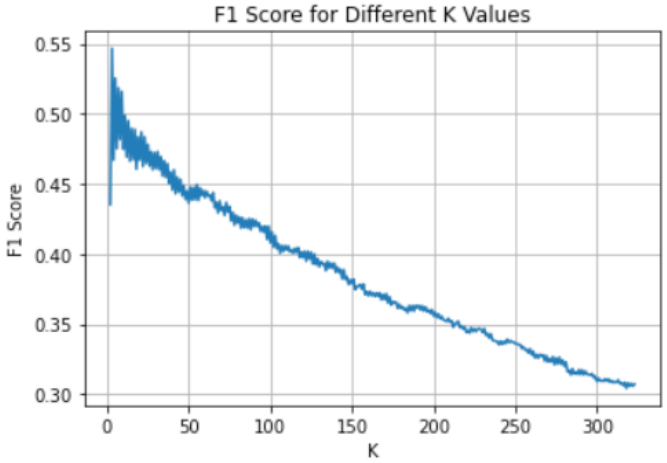


Fig. 22. F1 score v/s K value

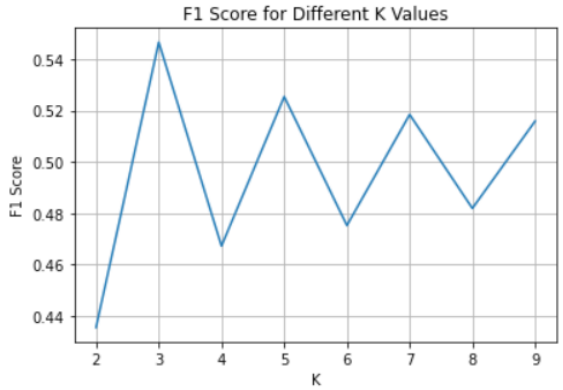


Fig. 23. max F1 score at K = 3

Methodology – Model testing and Evaluation

Metrics for Evaluation

- **F1 score** is a measure of a test's accuracy and is the harmonic mean of precision and recall. It considers both precision and recall to calculate the score, and provides a more balanced evaluation than just considering either precision or recall alone.

- **Precision** is the ratio of true positives (TP) to the sum of true positives and false positives (TP + FP). Precision, measures the proportion of positive cases that were correctly identified by the model out of all the cases that the model predicted as positive. It answers the question, "Out of all the predicted positive cases, how many were actually positive?"
- **Recall** is the ratio of true positives (TP) to the sum of true positives and false negatives (TP + FN). Recall, also known as sensitivity or true positive rate, measures the proportion of actual positive cases that were correctly identified by the model. It answers the question, "Out of all the actual positive cases, how many did the model correctly identify as positive?"

Methodology – Model testing and Evaluation

- A **confusion matrix** is a table used to evaluate the performance of a classification model by comparing the predicted labels with the actual labels. It displays the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each class.

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Fig. 24. Confusion Matrix for 2 class problem

Methodology – Model testing and Evaluation

SVM – without dimensionality reduction

Metric	Value
F1 score	0.6559
Precision	0.7265
Recall	0.6045

Table 2. Evaluation metrics for SVM

CD	HYP	MI	NORM	STTC
220	27	28	36	33
42	142	28	9	33
59	9	63	26	36
61	13	9	223	4
34	14	14	17	72

Fig. 25. Confusion Matrix for SVM

SVM – with dimensionality reduction using PCA

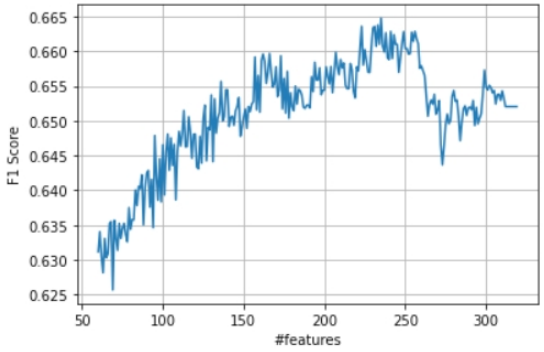


Fig. 26. F1 score v/s number of Features for SVM

Metric	Value
F1 score	0.6655
Precision	0.7455
Recall	0.6089

Table 3. Evaluation metrics for SVM after dimensionality reduction (235 features)

CD	HYP	MI	NORM	STTC
222	28	23	39	32
40	150	27	6	31
54	10	68	25	36
59	15	7	220	9
36	16	10	17	72

Fig. 27. Confusion Matrix for SVM after dimensionality reduction

Methodology – Model testing and Evaluation

KNN – without dimensionality reduction

Metric	Value
F1 score	0.5467
Precision	0.6749
Recall	0.4887

Table 4. Evaluation metrics for KNN

CD	HYP	MI	NORM	STTC
187	23	28	80	26
40	93	23	41	57
32	10	47	65	39
33	9	15	241	12
27	16	9	35	64

Fig. 28. Confusion Matrix for KNN

KNN – with dimensionality reduction using PCA

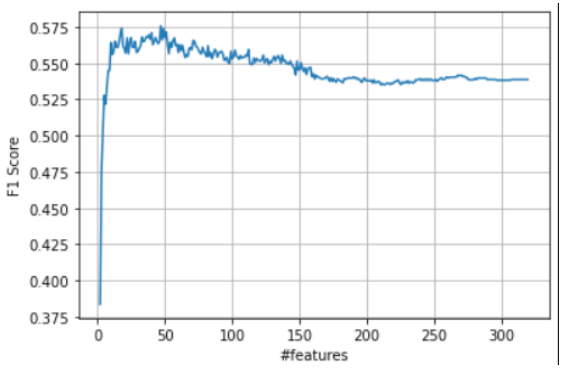


Fig. 29. F1 score v/s number of Features for KNN

Metric	Value
F1 score	0.5675
Precision	0.6629
Recall	0.5129

Table 5. Evaluation metrics for KNN after dimensionality reduction (47 features)

CD	HYP	MI	NORM	STTC
187	23	33	74	27
36	103	28	30	57
49	9	46	49	40
45	13	15	221	16
30	10	11	31	69

Fig. 30. Confusion Matrix for KNN after dimensionality reduction

Methodology – F1 score for different data split ratio

Train – Test split (% - %)	F1 score for SVM without PCA (%)	F1 score for SVM with PCA (%)	F1 score for KNN without PCA (%)	F1 score for KNN with PCA (%)
80 – 20	64.93	64.25	54.15	53.91
70 – 30	**65.49**	**65.59**	54.23	53.67
65 – 35	65.37	64.75	**54.67**	**54.28**
60 – 40	65.01	65.02	53.97	54.25
50 – 50	64.88	64.83	54.02	53.75

Table 6. F1 score comparison for different train and test split

Conclusion and Future Scope

- This project focuses on data processing, feature extraction, and machine learning model design and evaluation for ECG signal classification (PTB XL dataset).
- We explore two ML models, KNN and SVM. Both the models were trained and tested on the entire feature space (morphological features, statistical features and frequency domain features), as well as the reduced feature space using PCA. F1 score is the metric of evaluation.
- The SVM with dimensionality reduction using PCA (considering 235 features) has the best F1 score of 66.55% compared to other models.
- The drawback of SVM model is, it is suitable for a binary classification and performs one vs rest classification in a multiclass problem. The result can be improved by classifying the dataset into normal (NORM) vs abnormal (CD, HYP, STTC, MI)
- The future work will be use of additional features such as age, gender etc. for training and testing the models.
- Exploring different ML models such as Decision Tree, Random Forest, non-linear SVM etc.
- Using K-Fold cross validation technique for evaluation for better training and testing.

References

[1] *wikipedia.org. Electrocardiography. https://en.wikipedia.org/wiki/Electrocardiography.*

[2] *kaggle.com. PTB-XL ECG dataset. https://www.kaggle.com/datasets/khyeh0719/ptb-xl-dataset*

[3] *Wagner, P., Strodthoff, N., Bousseljot, R., Samek, W., & Schaeffter, T. (2022). PTB-XL, a large publicly available electrocardiography dataset (version 1.0.3). PhysioNet. https://doi.org/10.13026/kfzx-aw45.*

[4] *Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F.I., Samek, W., Schaeffter, T. (2020), PTB-XL: A Large Publicly Available ECG Dataset. Scientific Data. https://doi.org/10.1038/s41597-020-0495-6.*

[5] *Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.*

[4] *Z. Liu, J. Wang and B. Liu, "ECG Signal Denoising Based on Morphological Filtering," 2011 5th International Conference on Bioinformatics and Biomedical Engineering, Wuhan, China, 2011, pp. 1-4, doi: 10.1109/icbbe.2011.5780239.*

[5] *Rafael Dias Ribeiro de Almeida, Rômulo Martins Ponte. Design of a minimum power solution of ECG monitoring for edge applications. Semester Project in Sensor Signal Processing, 2021*

[6] *Piekarski, Krzysztof & Tadejko, Pawel & Rakowski, Waldemar. (2006). Properties of Morphological Operators Applied to Analysis of ECG Signals. 10.1007/978-0-387-36503-9_26.*

References

[7] *Taouli SA, Bereksi-Reguig F (2019) The QRS complex detection using morphological filtering. Biomed Sci Eng 5(1): 001-006. DOI: 10.17352/abse.000011*

[8] *Shankar, M. Gowri, and C. Ganesh Babu. "An exploration of ECG signal feature selection and classification using machine learning techniques." Int. J. Innovative Technol. Exploring Eng. Regul 9.3 (2020): 797-804.*

[9] *Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., & Chen, S. A. (2021). NeuroKit2: A Python toolbox for neurophysiological signal processing. Behavior Research Methods, 53(4), 1689–1696. https://doi.org/10.3758/s13428-020-01516-y*

[10] *Pham, T., Lau, Z. J., Chen, S. H. A., & Makowski, D. (2021). Heart Rate Variability in Psychology: A Review of HRV Indices and an Analysis Tutorial. Sensors, 21(12), 3998. https://doi:10.3390/s21123998*

[11] *Pan, J. and Tompkins, W., 1985. A Real-Time QRS Detection Algorithm. IEEE Transactions on Biomedical Engineering, BME-32(3), pp.230-236.*

[12] *S. K. Shrikanth Rao, M. H. Kolekar and R. J. Martis, "Frequency Domain Features Based Atrial Fibrillation Detection Using Machine Learning And Deep Learning Approach," 2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, 2021, pp. 1-6, doi: 10.1109/CONECCT52877.2021.9622533.*

References

[13] *G. T. Reddy et al., "Analysis of Dimensionality Reduction Techniques on Big Data," in IEEE Access, vol. 8, pp. 54776-54788, 2020, doi: 10.1109/ACCESS.2020.2980942.*

[14] *Rushikesh Pupale. Support Vector Machines(SVM) — An Overview. https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989*

[15] *Ajay Yadav. SUPPORT VECTOR MACHINES(SVM). https://medium.com/towards-data-science/support-vector-machines-svm-c9ef22815589*

[16] *Renu Khandelwal. K-Nearest Neighbors(KNN). https://medium.com/datadriveninvestor/k-nearest-neighbors-knn-7b4bd0128da7*

[17] *scikit-learn.org. GridSearchCV.*

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Code

Data_visualize_PTB_xl.ipynb – contains the code to load the dataset and visualize the data

main.ipynb – contains the code to load the dataset, feature extraction, PCA dimensionality reduction and evaluation of SVM and KNN

Model_train_inference.ipynb – contains the code for SVM and KNN model training and evaluation for different dataset splits.