Classification of ECG signals using Machine Learning algorithms

Gaurav Bajpai
MEng. Electrical and Computer Engineering
Toronto Metropolitan University
Toronto, Canada
gaurav.bajpai@ryerson.ca

***Abstract*—Machine learning algorithms can help in automatically detect patterns in data, drive data based decisions and impact the way clinicians make diagnosis, prognosis, and treatment of diseases. In this report Adaptive filter weights, mean, and variance of ECG signals are used as features, Logistic Regression, k-nearest neighbors(KNN) and AdaBoost classifiers are used to classify the ECG signals into normal and abnormal signals. Principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) are applied to visualize the two classes of data in two-dimensional map. Metrics accuracy, precision, recall and F1 score are used to identify the best algorithm for long term and short term signals. [1].**

***Index Terms*—ECG, Adaptive filters, machine learning, wearables**

I. Introduction

With technological innovations of Internet of Wearable Things (IoWT), a broader area of Internet of Things (IoT) and digital health has helped healthcare professional continuously and remotely monitor their patients' physiological conditions which earlier could only be monitored in a clinical setting thus providing adequate data to clinicians for earlier diagnosis, prognosis and treatment.The global wearables market is expected to grow from USD 14.93 billion in 2020 to USD 30.88 billion in 2025. The below figure1 shows different sensors used to detect bio-signals that could be used to diagnose diseases, mental health, and wellness.

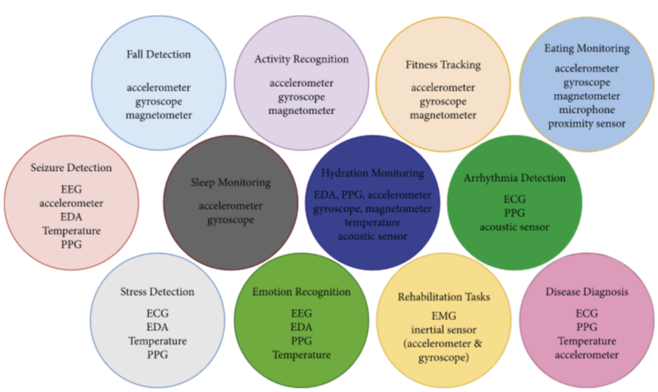

Fig. 1. Different sensors and bio-signals [2]

The 12-lead ECG Analysis is useful in the diagnosis and treatment of patients with acute myocardial infarction (AMI). ECG Analysis is also useful in the interpretation and documentation of other transient cardiac arrhythmias that may occur. When used in the prehospital setting, the 12-lead analysis results can be of assistance in diagnosis and treatment decisions once the patient has arrived in the hospital emergency department [3].

ECG analysis is done by placing electrodes on different locations on the surface of the body and recorded as the output waveforms. 12-lead ECG is the standard test both in non-emergency and emergency department along sides various other pathological examinations used by physicians for diagnosis, prognosis, monitoring, and treatment of both acute or chronic heart diseases.In a smartwatch, the ECG usually takes around 30 seconds to complete and data collected can be sent to a smartphone app, where it can be stored, viewed and forwarded to a doctor for remote monitoring. ECG is a short-term signal and may not detect infrequent arrhythmia. A Holter monitor helps clinicians to spot irregular heart rhythms that an ECG missed and is worn for 1 to 2 days to continuously records all ECG signals. These recordings are fed into a holter software and then analyzed for different heart-rate variability parameters to diagnose arrhythmia. Apple's smartwatch is an example of a FDA-cleared wearable device integrated with Convolutional Neural Network (CNN), a deep learning algorithm which can classify normal and abnormal sinus rhythms and capable of alerting its user to abnormal heart rhythms or arrhythmia. The existing state-of-the-art machine learning algorithms can be applied to medical data such as X-ray, MRI images or to bio-signals such as electroencephalogram (EEG) or electrocardiogram (ECG) to get classify signals into binary or multi-class classification to get meaningful insights of the patient health and get accurate diagnosis. In this report long-term and short-term ECG signals are classified using Logistic Regression, KNN and Adaboost algorithm (ensembles decision trees) into two classes normal and abnormal signals [4].

II. Materials and Methods

The ECG signals used in this report is obtained from physionet.org from two databases, MIT- BIH Arrythmia

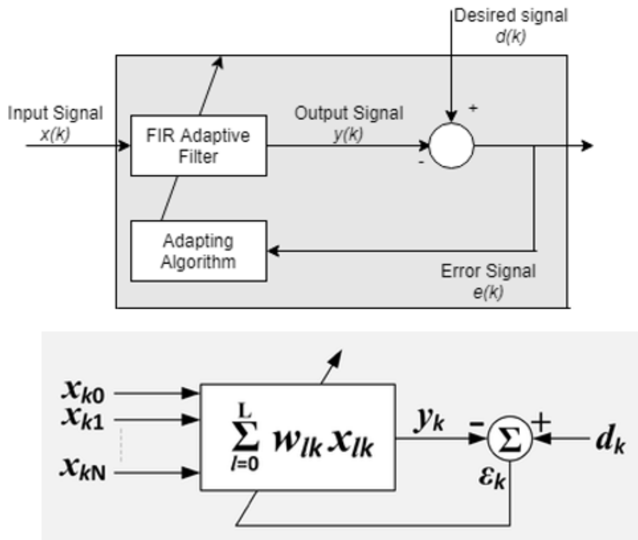

Fig. 2. Adaptive Filter block diagram [2]

Database and MIT-BIH Normal Sinus Rhythm Database. These databases have long-term ECG recordings over 30 minutes long. ECG recordings of 5 subjects 111, 115, 121, 209, 234 are taken from MIT- BIH Arrythmia Database are considered abnormal signals and labelled as 'zero'. These ECG signals have 650,000 samples and sampling rate of 360 Hz. ECG recordings of 5 subjects 16265, 16420, 16795, 18184, 19140 are from MIT-BIH Normal Sinus Rhythm Database are normal signals and labelled as 'one'. These ECG signals have 1,000,000 samples and sampling frequency of 128Hz. All the ECG signals have two signals modified limb lead II (MLII) (obtained by placing the electrodes on the chest) and a modified lead V1 (occasionally V2 or V5). MLII is used as the signal for the analysis in this report which records the changes likely to happen in ECG wave amplitudes during rest or exercise. A 60 Hz noise is added to this signal in MATLAB to make the signal noisy.

The block diagram of an adaptive filter is shown in figure [2]. It has an input signal x(k), output signal y(k), desired signal d(k) and error e(k) which is used by the adaptive algorithms to train the weights. The input signal x(k) in the analysis is the ECG signal and the desired signal d(k) is the signal plus the 60 Hz noise.

The algorithm used is Normalized Least Mean Squares(LMS).The adaptive algorithms work on minimizing the error, select the optimal filter coefficients and update the weights when new data arrives. The LMS is based on minimizing the error e(k). The adaptive filter trains weights to generate the approximate ECG signal. These filter weights are used as the features of the data-set given to the machine learning algorithms.

The LMS is implemented in Matrix form using and dsp.LMSFilter method to Normalized LMS from MATLAB software. The parameters for LMS algorithms are the number of weights and stepsize also called mu. The number of weights=20 and mu= 0.005 for all the LMS algorithms. [5], [6].

There are two datasets analyzed. The first dataset has 20 adaptive weights, mean of the signal and variance of the signal from the 10 original signals as features. This dataset consists of 10 rows with 22 features and a target column.The size of dataset1 is 10x23. For the second dataset dataset2, the 10 signals are divided into multiples of 50000 samples each. The normal sinus rhythm signals produce 100 rows and arrhythmia signals produce 65 rows and after data augmentation by copying few rows produce 105 rows. Dataset2 has 205 rows with 22 features and a target column (same as dataset1).

The datasets are imported into Python and Scikit-learn machine learning library is used to implement the algorithms. In dataset1, to find the linear correlation between the features, that is weights (w0 to w19) of the adaptive filter, the Pearson's correlation coefficients are calculated and the features that have a coefficient higher than 0.9, which indicate high correlation are dropped ('W10','W11', 'W12','W13', 'W14', 'W15', 'W16', 'W17', 'W18','W19', 'W6', 'W7', 'W8', 'W9','Sigmean'). The new dataset has a size of 10x7. For logistic regression using gridsearch calculated the regularization strength 'C': 0.001. Logistic regression is trained with 80% of the data and tested with 20% of the data. Then a 5-fold cross validation is done for Logistic regression with 80% of the data and tested with 20% of the data. Confusion matrix, accuracy, precision, recall and f1 score are calculated for simple train-test split and 5 fold cross-validation respectively for all the three classifiers with reduced datasets (one from dropping after calculating pearson's coefficients and from applying PCA). For KNN, the hyperparameter number of neighbors is calculated 'n-neighbors': 4. KNN is trained with 80% of the data and tested with 20% of the data. Then a 5-fold cross validation is done with 80% of the data and tested with 20% of the data. For Adaboost Algorithm optimal number of estimators or decision trees calculated using gridsearch is 10 'n-estimators': 10. Adaboost classifier is trained with 80% of the data and tested with 20% of the data. Then a 5-fold cross validation is done with 80% of the data and tested with 20% of the data. PCA algorithm is applied on dataset1 (with 10x22 size) to obtain 5 principal components, the last component is dropped, the size of resultant dataset is 10x4. Logistic Regression, KNN and Adaboost are applied on reduced dataset after PCA and each model trained is with 80% of the data and tested with 20% of the data. Then a 5-fold cross validation is done with 80% of the data and tested with 20% of the data. The first two principal components and two embedded features from t-SNE algorithm are used to produce two individual plots to visualize the separation of normal represented with '1' and abnormal represented with '0' in dataset1 [7].

For dataset2, size 205x22, Pearson correlation coefficient is applied to the features of dataset2 to find features with higher coefficient value 0.9 and above and those features are dropped. The resultant dataset2 has features

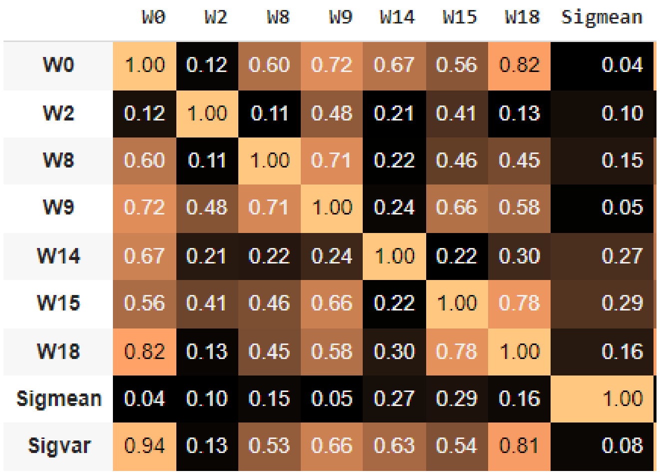

Fig. 3. Pearson correlation of dataset2

W0,W2,W8,W9,W14,W15,W18,and Sigmean; the size of dataset2 is 205x8. Optimal parameters are calculated using grid search for Logistic Regression, KNN and Adaboost, optimal parameters C, n-neighbors and n-estimators are calculated respectively similar to dataset1. Each of these classifiers is trained with 80% of the data and tested with 20% of the data. Then a 5-fold cross validation is done with 80% of the data and tested with 20% of the data. PCA algorithm is applied on dataset2 (with 205x22 size) to obtain 5 principal components, the last component is dropped, the size of the resultant dataset is 205x4. Then Logistic Regression, KNN and Adaboost are applied on reduced dataset after PCA and each model trained is with 80% of the data and tested with 20% of the data. Then a 5-fold cross validation is done with 80% of the data and tested with 20% of the data. The first two principal components and two embedded features from t-SNE algorithm are used to produce two individual plots to visualize the separation of normal represented with '1' and abnormal represented with '0' in dataset2.

$$Precision = \frac{TruePositive}{PredictedPositive} \tag{1}$$

$$Recall = \frac{TruePositive}{ActualPositive} \tag{2}$$

$$F1score = \frac{2 * Precision * Recall}{Precision + Precision} \tag{3}$$

III. Discussion and Results

Pearson correlation matrix of dataset2 is shown in figure below

Plot using principal components, PC1 and PC2 for dataset1 is shown in below figure normal represented with 1 and abnormal represented with 0

t-SNE plot with two components for dataset1 is shown in below figure normal represented with 1 and abnormal represented with 0

Plot using principal components, PC1 and PC2 for dataset2 is shown in below figure normal represented with 1 and abnormal represented with 0

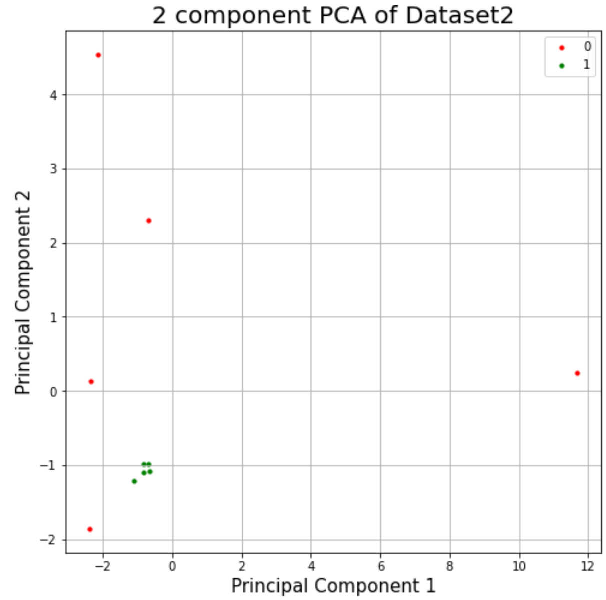

Fig. 4. PC1 and PC2 to visualize data-points normal represented with 1 and abnormal represented with 0

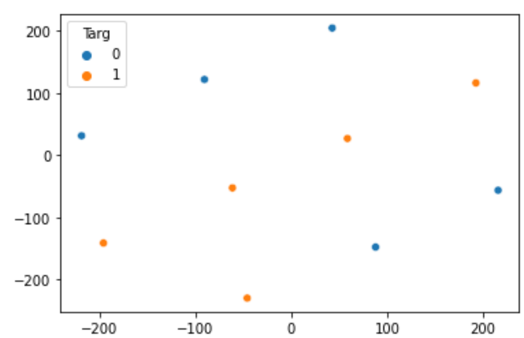

Fig. 5. t-SNE plot to visualize data-points normal represented with 1 and abnormal represented with 0

t-SNE plot with two components for dataset2 is shown in below figure normal represented with 1 and abnormal represented with 0

TABLE I
Logistic Regression for Dataset1

Metric	LR	5-fold LR	LR-PCA	5-fold LR-PCA
Test Accuracy	1.0	0.8	1.0	0.9
Test Precision	1.0	0.7	1	0.85
Test Recall	1	0.8	1	0.9
Test F1score	1	0.74	1	0.87

The original ECG is non-linear, non-stationary, non-guassian and non-shorterm. The split ECG signal is non-linear, non-stationary, non-guassian and non-shorterm. The original ECG doesnot address volume, variety, addresses velocity- high processing speed and veracity as it is performed in controlled environment. The split ECG signal address volume, not va-

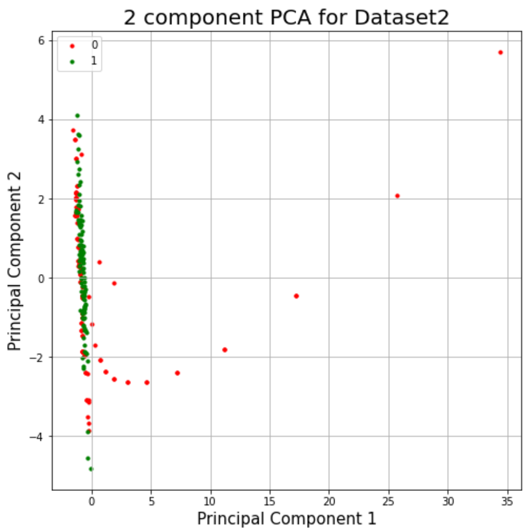

Fig. 6. PC1 and PC2 to visualize data-points normal represented with 1 and abnormal represented with 0

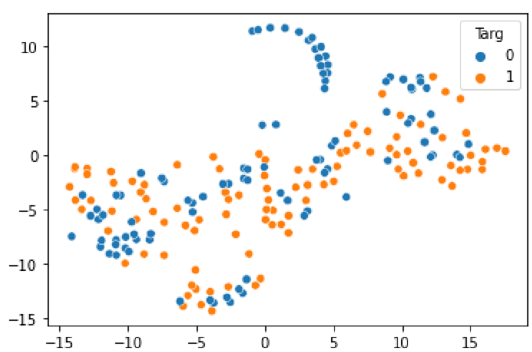

Fig. 7. t-SNE plot to visualize data-points normal represented with 1 and abnormal represented with 0

TABLE II
LOGISTIC REGRESSION FOR DATASET2

Metric	LR	5-fold LR	LR-PCA	5-fold LR-PCA
Test Accuracy	0.78	0.57	0.68	0.51
Test Precision	0.8	0.59	0.69	0.26
Test Recall	0.78	0.57	0.68	0.5
Test F1score	0.78	0.57	0.68	0.34

TABLE III
KNN FOR DATASET1

Metric	KNN	5-fold KNN	KNN-PCA	5-fold KNN-PCA
Test Accuracy	1.0	0.8	1.0	1.0
Test Precision	1.0	0.7	1	1
Test Recall	1	0.8	1	1
Test F1score	1	0.74	1	1

TABLE IV
KNN FOR DATASET2

Metric	KNN	5-fold KNN	KNN-PCA	5-fold KNN-PCA
Test Accuracy	0.8	0.82	0.85	0.76
Test Precision	0.81	0.86	0.86	0.79
Test Recall	0.8	0.82	0.85	0.76
Test F1score	0.8	0.83	0.85	0.77

TABLE V
ADABOOST FOR DATASET1

Metric	Adaboost	5-fold Ada.	Ada.-PCA	5-fold Ada-PCA
Test Accuracy	1.0	0.8	1.0	0.6
Test Precision	1.0	0.7	1	0.4
Test Recall	1	0.8	1	0.6
Test F1score	1	0.74	1	0.48

TABLE VI
ADABOOST FOR DATASET2

Metric	Adaboost	5-fold Ada.	Ada.-PCA	5-fold Ada-PCA
Test Accuracy	0.95	0.88	0.88	0.88
Test Precision	0.95	0.89	0.88	0.88
Test Recall	0.95	0.88	0.88	0.88
Test F1score	0.95	0.88	0.88	0.88

riety, addresses velocity- high processing speed and veracity as it is performed in controlled environment. Among the algorithms F1-score metric for dataset1 the KNN and then Logistic Regression was better in all the four combinations.Among the algorithms for dataset2 F1-score metric for Adaboost and then KNN was better in all the four combinations. The PCA and t-SNE plots for original dataset clearly separates normal and abnormal dataset. For the split dataset t-SNE separates the dataset better than PCA visually. [3] [5].

REFERENCES

[1] L. V. R. Kumari, A. jyothi Sabavat, and Y. P. Sai, "Performance evaluation of adaptive filtering algorithms for denoising the ecg signal," *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 1–5, 2021.

[2] K. Elangovan, "Comparative study on the channel estimation for ofdm system using lms, nlms and rls algorithms," *International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012)*, pp. 359–363, 2012.

[3] M. E. Hanine, E. H. Abdelmounim, R. Haddadi, and A. Belaguid, "Real time emg noise cancellation from ecg signals using adaptive filtering," in *ICCWCS'17*, 2017.

[4] S. R. Tithi, A. Aktar, and F. Aleem, "Machine learning approach for ecg analysis and predicting different heart diseases," 2018.

[5] A. P., R. Arabelli, and R. K, "Noise removal of ecg signals with adaptive filtering," *Materials Today: Proceedings*, 2021.

[6] M. K. Islam, A. F. Haque, G. Tangim, T. Ahammad, and M. R. H. Khondokar, "Study and analysis of ecg signal using matlab & labview as effective tools," *International Journal of Computer and Electrical Engineering*, pp. 404–408, 2012.

[7] A. Platzer, "Visualization of snps with t-sne," *PLoS ONE*, vol. 8, 2013.