International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 08, 2020
ISSN: 1475-7192

Analysis on Heart Arrhythmia prediction and Classification

[1]Raima Mathew, [2]Ganesh Madarasu, [3]G Manju

Abstract--*Machine Learning is an integral part of Artificial Intelligence, is a science of statistical models and their algorithms. It is used to train a model to perform a real-world task without explicit commands. Heart Arrhythmia is a life-threatening disease dealing with an irregular heartbeat. ECG signals are the most accurate measure to find out the functionality of the cardiovascular-system. We put forward a solution to solve the difficulty of choosing among the various state-of-art models for heart arrhythmia prediction. This helps the hospitals with less experienced doctors to more accurately predict the disease at a lower cost. The Convolutional Neural Network model we built is more accurate than the existing models. We have also analyzed with other machine learning models such as Logistic Regression, Random Forest and XGBoost for the hospitals with lesser computation power.*

Key words--*Machine Learning, Artificial Intelligence, ECG, Heart Arrhythmia, Convolutional Neural Network, XGBoost*

I. INTRODUCTION

According to a survey conducted by the World Health Organization (WHO) in 2016, about 17.9 million people suffered from cardiovascular disease based deaths. These counted more than one-fourth of all deaths of which most are due to heart attack or stroke. Even with this large number of deaths, the detection of heart diseases at early stages is almost impossible. With the juncture of medicine and machine learning, it is now possible to detect heart diseases at an early stage.

Machine Learning is making computers respond automatically without being told explicitly. In this process of training the computer, a model is built to perform the operation. The input data is given to the model and it performs statistical analysis to produce the output. The operation can vary according to the application.

The heart sends out rhythmic electric impulses for the proper functioning of the heart. A disparity in the rhythm causes heart arrhythmia. If the heartbeats as fast as - above hundred beats per minute, it is called "tachycardia". If the heartbeats are too slow- below sixty beats per minute, it is called bradycardia. The four types of heart arrhythmia are - supra-ventricular tachycardia, extra beats, ventricular arrhythmia, "bradyarrhythmia".

II. STATE OF THE ART

The state of art models of current research on the work is explained in the following column. We have conducted a survey for few papers that are the most relevant to our problem statement.

[1]*Department of Computer Science, SRM IST, Kattankulatur, Chennai, India, rm9700@srmist.edu.in*
[2]*Department of Computer Science, SRM IST, Kattankulatur, Chennai, India, gm6054@srmist.edu.in*
[3]*Assistant Professor, Department of Computer Science, SRM IST, Kattankulatur, Chennai, India, manjug@srmist.edu.in*

DOI: 10.37200/IJPR/V24I8/PR280112
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

In this paper, the cardiac arrhythmia is predicted using deep convolutional neural networks. The five types of arrhythmia are predicted and the acquired knowledge is transferred to the task of classification of myocardial infarction. In this method, 1-D convolution through time is applied in every layer of the architecture, kernels of size 5.

In the task of predicting Myocardial infarction, the weights of all the layers except the last two are frozen. Only the last two network layers are trained to perform the classification. The result attained is visualized in t-SNE (t-Distributed Stochastic Neighbor Embedding) to showcase the effectiveness of the approach.

This paper is based on detecting the prediction of heart diseases using "Support Vector Machine (SVM)". The learning and classification of the SVM are carried out by two modules: a training module and classification module for different heart conditions. The features are extracted in this model using "BIOPAC Acknowledge" software in order to model the diseases. The model is trained using the ECG features. The resultant classification is if the heart condition is normal or abnormal. If the condition is abnormal, the conclusion drawn is that the person suffers from cardiac arrhythmia.

The purpose of this paper is to detect cardiac arrhythmia from the clustering and regression approach. The clustering is carried out by "DBSCAN (Density based spatial Clustering of Application with Noise)" and the "Regression" is carried out by multiclass logistic regression. The dataset is divided in DBSCAN algorithm into sections that have instances that belong to more than one class. After DBSCAN is carried out, the logistic regression is applied to the chosen clusters in order to produce theta value. During testing, the class with the highest probability is considered to be the type of cardiac disease.

In this paper, the prediction of arrhythmia is carried out to predict one among the 16 subclasses. For this prediction, "Support Vector Machine (SVM)" methods are utilized such as "one-against-one (OAO), one-against-all (OAA), the error-correction-code (ECC)". The evaluation is performed using accuracy, kappa statistics, and root mean square. A large number of features in the dataset is reduced by feature selection technique known as the "wrapper method". This method is developed around "random forest algorithm". The features that have binary values as variables are removed as their contribution to classification is less. The SVM classifier outperformed the other machine learning models with an accuracy of 81 percent.

This paper focuses on predicting the point of source of cardiac arrhythmia. This is carried out by placing a reference catheter in the heart and another diagnostic catheter moves around the heart recording data. The prediction occurs dynamically as the data is being collected. The optimization used in this paper is Douglas-Rachford splitting. Since the second catheter is collecting data dynamically, it is known to have many noisy data. Hence potential outliers are removed when solving the optimization process. The outliers are identified when it is consistent and difficult to fit. About ten observations are considered of the patient from the diagnostic catheter. The initial ten observations are randomly found by the "Monte Carlo" sampling. Using this paper, the point of source of the disease is suggested with minimal information.

DOI: 10.37200/IJPR/V24I8/PR280112
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

In this paper the aim to classify normal and abnormal heart arrhythmia using a combination of Elephant Herding Optimization and Support Vector Machine. Their approach includes four parts such as preprocessing, "ECG feature extraction, feature selection and optimization, and classification". This is followed by validation. Noise is removed using modified pan Tompkins algorithms. Using bandpass filter and derivative integrator, the peaks and outliers are identified. In this method, more features are obtained using an improved feature extraction algorithm. The optimization algorithm is swarm-based algorithm. The behavior of a group of swarms are identified and their behavior with each other and with the environment is noted. This model's performance is measured using accuracy, sensitivity, precision, specificity, F-measure.

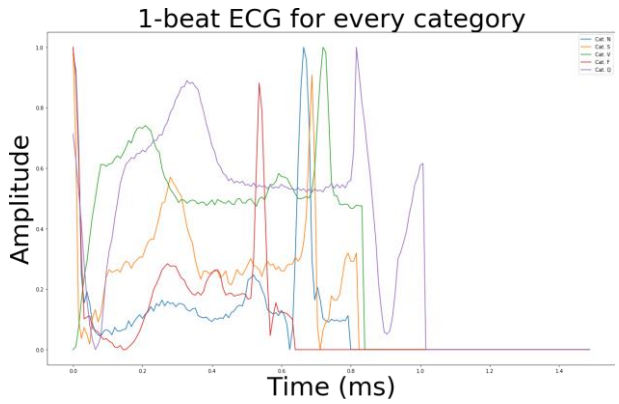

Figure 1: Signals of each category

In this paper, the focus is on anomaly detection. It is carried out in two phases. In the first phase, "random forest ensemble and recursive feature elimination method" is used. The groups of important features are then clustered using k-median algorithm. Towards the end of every stage, the data is clustered into groups that are not related. This makes the training phase easier. The oversampling PCA is then carried out. It identifies the main peaks from the data. The outliers ranks provided to data instances and after completion of the oversampling principle component analysis process, the point that is considered to be suspected outliers are filtered out.

This journal focuses on the cardiovascular disease being at the mirror of science. The paper is to show the pattern of statistical activities in the domain of cardiovascular disease that had been carried out for a period of 10 years i.e from 2001-2010. There were numerous publications regarding the field of cardiovascular heart disease (CVD), mainly in North America and Western Europe. This result confirms the increasing amount of interest in this particular field. The World Health Organization (WHO) has confirmed that CVDs are the chief and the most prime reason for the deaths and disabilities around the world. A survey conducted resulted to the prediction that about 17.3 million people have died because of heart related diseases. A pattern that has been observed is that the deaths mostly occur in the poor and middle-income countries.

III. PROPOSED WORK

Heart arrhythmia is a very lethal disease when remained undetected and untreated. The manual analysis of detection of heart arrhythmia is very tiresome and requires a lot of experience along with knowledge. This cannot be

DOI: 10.37200/IJPR/V24I8/PR280112
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

carried out in small clinics and dispensaries where there is a scarcity of well-educated doctors. The existing automatic prediction and detection machine is about 3,60,000 rupees and hence is not affordable by all the customers and hospitals. With our analysis, the most optimum method to perform the detection will be discovered using which the disease can be predicted automatically in their early stages. This model can be attached to a pocket ECG monitor and the analysis can be done in a pocket friendly manner. The cost of a pocket ECG monitor is about 4500 rupees and hence could be afforded by common dispensaries. [1]The CNN created was not as accurate as our model, the accuracy is a crucial factor because it is the detection of a lethal disease.

IV. DATASET

In this paper, the dataset used is PhysioNet MIT-BIH dataset. The dataset is acquired by Boston's Beth Israel Hospital laboratories and MIT research team. The MIT-BIH consists of 48 half an hour portions of ECG signals collected from different patients (47). All these heartbeats are sampled at the rate of 360Hz. All the beats are categorised by more than two cardiologists. The ECG signals are of five various categories of cardiac arrhythmia disease according to the "Association of Advancement of Medical Instrumentation (AAMI) EC57 standard".

V. IMPLEMENTATION

In this section, we describe the machine learning algorithms that has been used and compared.

Random Forest Algorithm

The random forest algorithm is an algorithm that considers many decision trees. The algorithm works flawlessly because the prediction of different trees is considered and the prediction of each tree is taken into account. The result is based on the majority of the outputs from the trees. The different trees are uncorrelated. The non-correlation is maintained by methods like bagging. It takes advantage of the fact that a small change in the input dataset causes a major difference in the model prediction. The feature randomness is used such that the features that are most different from each other are taken into account. This brings in more diversification. This algorithm is written in python for this paper.

Logistic Regression

The logistic regression is one of the most fundamental algorithms in machine learning. It performs classification either binary classification or multiclass classification. In this case, multiclass regression is used. If the coefficients are good enough, the result obtained will be close to 1. In logistic regression, multiple gradient descent functions can be used, the most common one used is sigmoid function. The output value is produced by combining the input value along with the coefficients and weights.

XGBoost Algorithm

XGBoost algorithm is a decision tree based algorithm that utilises gradient boosting algorithm. It handles missing values in the dataset, performs regularization to avoid overfitting. It carries out the tree pruning by parallel

DOI: 10.37200/IJPR/V24I8/PR280112
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

data processing. This algorithm is built to optimize the use of hardware resources in the process. It avoids overfitting through LASSO and Ridge regularization. The multiple forms of gradient boosting are supported by XGBoost are gradient boosting, stochastic boosting, regularized gradient boosting.

Convolutional Neural Network

Using a "Convolutional neural network (CNN)", the input is taken and weights are assigned to the neurons which assign their importance in feature selection. The preprocessing required for CNN is comparatively lesser than the other classification algorithms. The input is a pixel of values and when it is trained enough, it can be used to perform complex classification. We preprocessed the dataset using One Hot Encoder from sk-learn that creates a binary column for every category and returns a sparse matrix. There are 6 kernels created of size 5, at an appropriate learning rate with multiple fully connected layers before the final softmax layer. The validation set was used along with the training and test set in order to avoid overfitting and improve the performance.

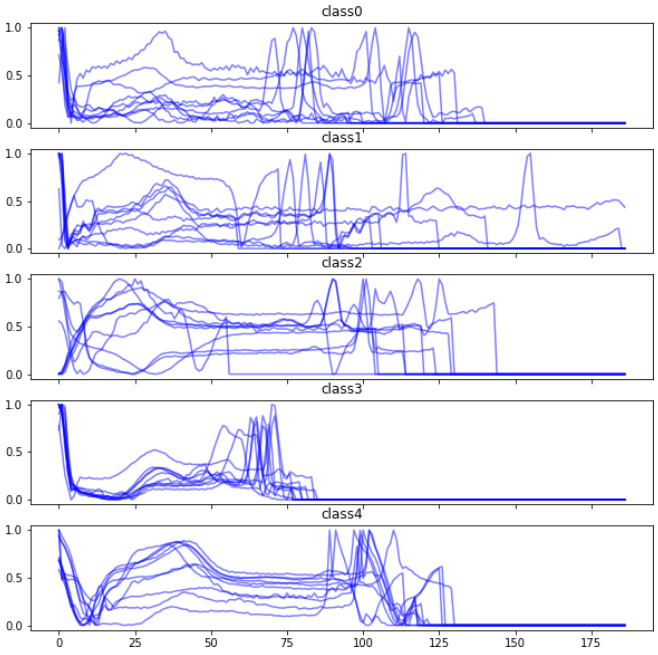

Figure 2: Signals before preprocessing

VI. PREPROCESSING

In the machine learning model building process, the most important and tiring part is the preprocessing of the data. Data collected initially contains outliers, noisy data, etc. These causes the data to be affected and interpreted in the wrong way. Hence removal of these factors is highly important.

DOI: 10.37200/IJPR/V24I8/PR280112
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

During preprocessing, there is a check for the existence of null values, empty values, extreme values. These conditions are removed or altered in this process. Along with this, the values are also normalized. For example, the values are brought to the same range so that it is easier to calculate and perform prediction.

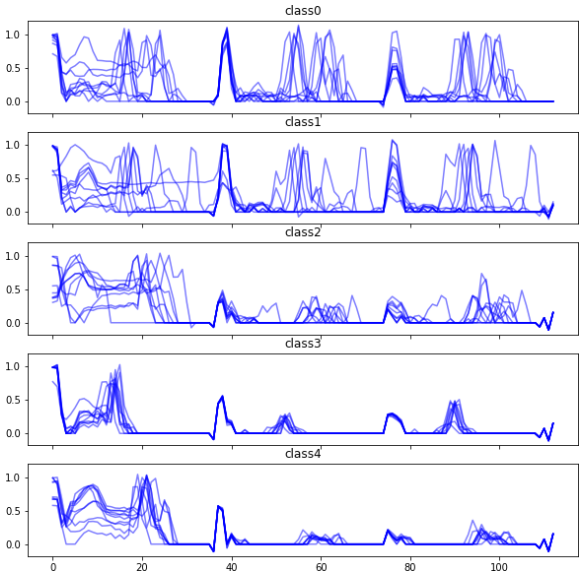

Figure 3: Signals after preprocessing

The extreme values are not appreciated in the statistical field of work. And hence the values are always smoothed and made easier to perform calculations on. The figure 3 shows the difference between the original data and the smooth data. There is a stark difference in the cost of computation due to this one factor. Data Augmentation is also carried out because the dataset contained more number of cases with normal heartbeat. This would make the model to be biased towards normal heartbeat over the other cases. Hence the data for the other classes are augmented.

Table 1: Table of performance analysis of this paper

S. No	Model Name	Accuracy (%)
1.	Logistic Regression	88
2.	Random Forest	97
3.	XGBoost	96
4.	CNN	99

VII. RESULTS DISCUSSION

The accuracy is one of the major factor in evaluating the performance of any model. In Logistic Regression at the cost of 28 seconds, we achieve an accuracy of 88 percentage. Even though this accuracy is low, it can be used in hospitals with very less computation. In random forest algorithm, the accuracy was 97 percentage with a loss of 0.26. In XGBoost algorithm, the accuracy is similar to the random forest algorithm but the log loss is 0.13. In convolutional neural network, the accuracy obtained was 99 percentage with a log loss of 0.08. This is the best

DOI: 10.37200/IJPR/V24I8/PR280112
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

way to predict the cardiac arrhythmia if there is availability of high computation. The training was carried out in Amazon Web Services and Google Colab.

VIII. FUTURE WORK

The current work can be extended by using different deep learning algorithms like recurrent neural network, convolutional recurrent neural network. This might perform better than the proposed work. And more datasets of premature arrhythmia and fusion of normal and ventricular arrhythmia can provide better performance as they are very less in number in the used dataset.

IX. CONCLUSION

This papers aims to provide a simple and one-step solution to the question of best algorithm to be used in prediction of Heart arrhythmia. Using this, the prediction of the disease can be performed in early stages when it can be cured and be prevented from causing more damage to the patient. The current work can be extended by using different deep learning algorithms like recurrent neural network, convolutional recurrent neural network. Using this information, even inexperienced doctors or doctor assistants can suggest the patients regarding the probability of a heart disease in them.

REFERENCES

1. ECG Heartbeat Classification: A Deep Transferable Representation. Mohammad Kachuee, Shayan Fazeli, Majid Sarrafzadeh - University of California, Los Angeles (UCLA)
2. An approach of cardiac disease prediction by analyzing ECG signal. Published in 2016 3rd International Conference on Electrical Engineer- ing and Information Communication Technology (ICEEICT) by
3. Prediction of cardiac arrhythmia type using clustering and regression approach (P-CA-CRA) published by IEEE. Authors are Prathibhamol Cp ; Anjana Suresh ; Gopika Suresh. Published in 2017 International Conference on Advances in Computing, Communications and Informat- ics (ICACCI)
4. Multiclass Classification of Cardiac Arrhythmia Using Improved Feature Selection and SVM Invariants. Anam Mustaqeem, Syed Muhammad Anwar and Muahammad Majid - University of Engineering and Tech- nology, Taxila, Pakistan.
5. Computer Assisted Localization of a Heart Arrhythmia published in 2018 by Chris Vogl, Peng Zheng, Stephen P. Seslar, Aleksandr Y.
6. Aravkin. https://arxiv.org/abs/1807.03091
7. Combining Support Vector Machine and Elephant Herding Optimization for Cardiac Arrhythmias published by Aboul Ella Hassanien1, Moataz Kilany2, and Essam H. Houssein from Computer Science Department, Minia University, Egypt.
8. Attribute Reduction based Anomaly Detection Scheme by Clustering Dependent Oversampling PCA by Asha Ashok, Smitha S, Kavya Krishna MH of Computer Science and Engineering Amrita School of Engineering, Amritapuri. Published in 2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, 2016.
9. Cardiovascular diseases in the mirror of science by MohammadHos- sein Biglu, Mostafa Ghavami,and Sahar Biglu. Published online on 10.15171/jcvtr.2016.32
10. Cardiac Arrhythmia classification by Multi-layer perceptron and Convolutional neural networks by Shalin Savalia and Vahid Emamian
11. Support Vector Machines for Anomaly detection by Xueqin Zhang, Chunhua Gu, and Jiajun Lin
12. Masoud, Shiravand, Rezapour Maryam, Rezapour Sadegh, Mardani Mahnaz, and Bahmani Mahmoud. "A Review of Medicinal Plants Affecting Exercise and Physical Health Factors in Athletes." Journal of Complementary Medicine Research 10 (2019), 212-225. doi:10.5455/jcmr.20190821081803

DOI: 10.37200/IJPR/V24I8/PR280112
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

13. Khode, V., Sindhur, J., Kanbur, D., Ruikar, K., Nallulwar, S. Mean platelet volume and other platelet volume indices in patients with stable coronary artery disease and acute myocardial infarction: A case control study(2012) Journal of Cardiovascular Disease Research, 3 (4), pp. 272-275. DOI: 10.4103/0975-3583.102694

DOI: 10.37200/IJPR/V24I8/PR280112
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020