Engineering and Technology Journal for Research and Innovation (ETJRI)
ISSN 2581-8678, Volume IV, Issue I, Jan 2022

Detection of Parkinson Disease UsingMachine Learning

[1]G.Priyadharshini, [2]T.Gowtham, [3]M.Harshavardhan Bhoopathi, [4]M.Reshma, [5]V.Tamilarasi, [6]P.Nandhini
[1] *Assistant Professor,* [2,3,4,5,6] *Under Graduate Students, Department of Computer Science and Engineering, Coimbatore Institute of Technology ,Coimbatore ,India*

***Abstract*-Biomarkers derived from human voice can offer in-sight into neurological disorders, such as Parkinson's disease (PD), because of their underlying cognitive and neuromuscular function. Historically, PD has been difficult to quantity and doctors have tended to focus on some symptoms while ignoring others, relying primarily on subjective rating scales. With advancements in technology and the prevalence of audio collecting devices in daily lives, reliable models that can translate this audio data into a diagnostic tool for healthcare professionals would potentially provide diagnoses that are cheaper and more accurate. This project provides evidence to validate this concept here using a voice dataset collected from people with and without PD. This paper explores the effectiveness of using extreme gradient boosting algorithms, such as XG boost algorithm, to accurately diagnose individuals with the diseaseResearchers have found that this term is a basic term for a PD More recently, Learning Machines (MLs) have helped solve computer vision problems, process natural languages, recognize Parkinson's speech using machine learning tools and provide a better understanding of the PD database in the current decade. Parkinson's voice database is available at UCI machine storage in the center of learning equipment and intelligent programs. The major outcome of this project is to bring out the result with at most accuracy using XG boost Algorithm. The database contains the similar links show high variability in the Parkinson's disease database. The most importantly the comparison has been made to explicit the outcome result and to prove that the XG boost algorithm provide more accuracy comparing with other algorithms.**

***Keywords*- Parkinson's disease, speech symptoms, machine learning, segregation, feature selection**

I. INTRODUCTION

A clinical diagnosis of Parkinson's disease (PD) can be confirmed based on onneuro-pathologic and histo-pathologic procedures. Features. Consideration of independent clinica lstudies. The number of patients showing PD is required to investigate clinical, pathologic, and nosologic studies depending on the frequency of occurrence, features, and risk factors in patients. Neural Networks, DMneural, Regressionand Decision Trees were previously hired to calculate the effectiveness of classification diagnostic classifiers. PD causes speech impairment affecting speech, motor skills, and other functions such as behavior, emotions, hearing and thinking. Tele-monitoring diagnoses using voicere levance and the importance of relationship statistics and PD symbols. Numerous diagnoses and considerations of various clinical features provide a diagnosis of Parkinson's disease. The precision section (ACC), Kappa Error (KE) and Area Receiver Operating Characteristic (ROC) Curve (AUC) with two basic elements, namely KStar and IBk provide a pre-diagnostic PD diagnostic model. Medical biometrics play an important role in diagnosing disorders such as PD. The evaluation of Parkinson's (PD) clinical diagnostic procedures demonstrates computer efficiency and the effectiveness of a class-bound approach to distinguishing healthy subjects from those with PD. The accuracy obtained with possible neural network training using the Parkinson's disease database was predicted using WEKA 3 and MatLab v7. Diagnosis based on the neural network of medical diseases over the past decade shows great attention to the prediction of PD. Parkinson's disease (PD) is a neuro-pathological disorder that affects the functioning of the human body. This is the second most common neurological disease seen after Alzheimer's disease and it is estimated that more than one million people suffer from PD in North America alone. In 1817, the PD was called shaking palsy by Dr. James Parkinson [4]. Various studies have shown that this figure will increase in the elderly as it is most often seen in people over the age of 60. Parkinson's disease is characterized by the breakdown of certain groups of brain cells responsible for the production of neurotransmitters including dopamine, serotonin and acetylcholine. Loss of dopamine effect on symptoms such as anxiety, depression, weight loss and vision problems. Other symptoms that can be seen in people with Parkinson's disease are incontinence, inability to speak properly and trembling various studies have shown that 90% of people with PD have oral problems including

Engineering and Technology Journal for Research and Innovation (ETJRI)
ISSN 2581-8678, Volume IV, Issue I, Jan 2022

dysphonia, monotone and hypophonia. Therefore, speech impairment is considered the first sign of Parkinson's disease. The cause and treatment of PD is not yet known but the availability of various drugs provides a significant reduction in symptoms especially in the early stages, thereby improving the quality of the patient and reducing the cost of Pathology. Voice rate analysis is simple and non-invasive. Thus, PD voice measurement can be used. To test the progress of the PD, various voice tests are scheduled including continuous calls and speech recording. Tele-monitoring and tele-diagnosis systems have been widely used as these systems rely on inexpensive and easy-to-use speech signals. Therefore, in this paper, there is an attempt to test the best machine-based learning model for early detection of PD from sample name samples.

II. LITERATURE REVIEW

From time to time, several notable attempts have been made by various investigators to diagnose Parkinson's disease. The following is a brief review of some of the activities performed to detect Parkinson's disease from voice samples from studies. Max A. little et al suggested a novel process on the classification of articles in Parkinson's sick studies and the management of dysphonia. In their work, the introduction of voice entropy (PPE) is a new measure of dysphonia. Data were collected from 31 individuals (23 were PD patients and 8 were healthy subjects) meaning there were 195 calls in progress. Their approach consisted of three stages; feature calculation, processing and selection of final items I have separated. For the purpose of separation, use them line kernel vector (SVM) auxiliary machine. Their proposal model obtained 91.4% accuracy. To separate healthy studies from PD topics, Ipsita Bhattacharya et al [20] used a data mining tool known as weka. They used SVM, a visual machine that reads algorithm for the purpose of partitioning. Prior to the sections, data processing was performed on file data. Different kernel values are used to obtain the best results accurately using libSVM. The SVM direct channel produced excellent accuracy of 65.2174%, while the RBF kernel and polykernel SVM achieved 60 accuracy. 8696%. In another work, BE Sakar et al [12] suggested a model for classifying control subjects from PD Articles. In their study, data were collected from 40 subjects (20 were healthy subjects and 20 were PD Courses). For each article, 26 voice samples were recorded that included short sentences, words, numbers, and stable vowels. Separate, use SVM and neighboring k (k-NN). To ensure crossing, use Summarized Leave-One-Out (s-LOO) and Leave- One-Subject-Out (LOSO). Numbers 1, 3, 5 and 7 selected for k-NN and SVM, linear and RBF kernel were used. 82.50% accuracy was achieved by k-NN and 85% accuracy was reported using the SVM classifier. Achraf Benba et al aims to exclude people with PD from control studies. In their work, data with 34 vowels, were collected from 34 people out of 17 who were PD subjects. From each topic, 1 to 20 Mel-frequency cepstral coefficients (MFCC) were obtained. SVM with various kernel types used for partitioning. LOSO has been used as a means of ensuring crossing. Excellent accuracy of 91.17% was reported by linear kernel SVM in taking the top 12 MFCC coefficients. With the discovery of PD, different signal processing algorithms were compared with C.O Sakar et al. In their work, a new feature called the tunable Q-factor wavelet transform (TQWT) was introduced. The performance of TQWT has surpassed the processing techniques of high-quality speech signals used to exclude features in the detection of PD. In different feature subsets, different classifications are used and the classifiers predicting techniques are combined. It was found that MFCCs and TQWT received very high clarity and were therefore considered important. features in the problem of PD fragmentation. Also, the process of selecting the minimum demolition feature has at least been used as a data preparation step. The highest accuracy of 86% was reported by RBF kernel SVM in all feature subsets. Richa Mathur et al [23] suggested a method of predict a PD. They have used a weird tool to use algorithms to perform data processing, classification and outcome analysis on a given database. They used k-NN and Adaboost.M1, bagging, and MLP. It was observed that k-NN + Adaboost.M1 produced a very good separation accuracy of 91.28%. A.Yasar et al [24] used artificial neural networks to get Parkinson's disease. The database is taken from

Engineering and Technology Journal for Research and Innovation (ETJRI)
ISSN 2581-8678, Volume IV, Issue I, Jan 2022

the machine's UCI learning archive. Using the MATLAB tool, 45 properties were selected as input values and one subdivision product. Their proposed model was able to separate

III. METHODOLOGY

Parkinson's disease in its early stages using machine learning techniques is shown in figure1 .It contains the following steps Parkinson's voice database is obtained from UCI machine storage from the Center for Machine Learning and Intelligent Systems. Matrix column entries (attributes) reflect the following definition: Name - ASCII title name and recording number MDVP: Fo (Hz)- Basic voice frequency measurement MDVP: Fhi (Hz) - The most common voice frequency MDVP: Flo (Hz) - Basic frequency of low volume MDVP: Jitter (%), MDVP: Jitter (Abs), MDVP: RAP, MDVP: PPQ, Jitter: DDP - Several measures of variability in basic frequency MDVP: Shimmer, MDVP: Shimmer (dB), Shimmer: APQ3, Shimmer: APQ5, MDVP: APQ, Shimmer: DDA - A few steps to vary in scope NHR, HNR - Two steps for measuring sound in specific parts of the voice by voice. condition - Health condition of topic (single) - Parkinson's, (zero) – healthy RPDE, D2 - Two complex steps for diversity DFA - Strengthening period spread1, spread2, PPE - Unusual frequency fluctuations The data contains 195 instances and 24 symbols.

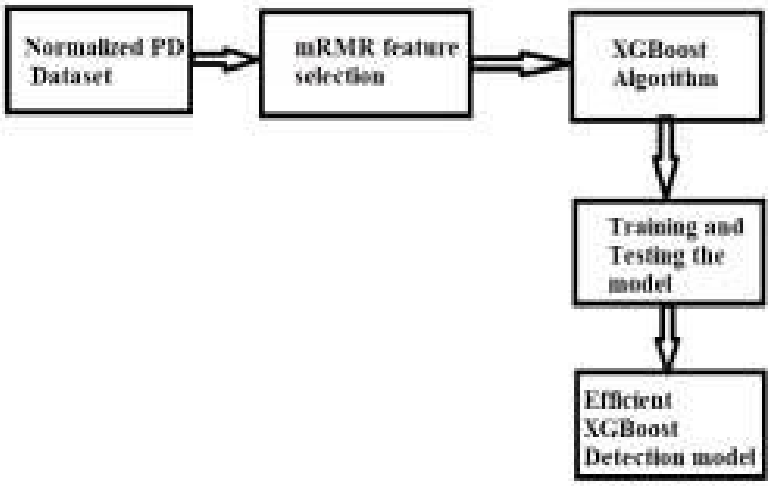

Fig. 1. Overview of the proposed framework name, MDVP: Fo (Hz), MDVP: Fhi (Hz), MDVP: FHz), MDVP: Jitter (%), MDVP: Jitter (Abs),MDVP: RAP, MDVP: PPQ, Jitter: DDP, MDVP: Shimmer , MDVP: Shimmer (dB), Shimmer: APQ3, Shimmer: APQ5, MDVP: APQ, Shimmer: DDA, NHR, HNR, status, RPDE, DFA, spread1, spread2, D2, PPE

A. PD Dataset

The first step in the process is data collection. Voice analysis, data are collected from UCI, a machine readable container containing voice data in both PD and healthy subjects. The database used has 195 instances and 24 symbols.

B. Data Pre-processing

This step is a combination of two individual processes, namely data processing and reduction of a feature or selection process described below.

C. Data Normalization

Data standardization is a method of preparing data that is commonly used in data sets while working with most machine learning algorithms. Changes column number values without losing any information. It is necessary to re-measure the values of a particular element in a particular range. In this function, the feature values in the selected database are standardized using the Min-Max measurement range in the range (0, 1) as they were in different distances.

Where X is a specific element represented by a column in the database, x I is the

Engineering and Technology Journal for Research and Innovation (ETJRI)
ISSN 2581-8678, Volume IV, Issue I, Jan 2022

value of this column where I am the number of objects in the

$$MinMax(X) = \frac{x_i - X_{min}}{X_{max} - X_{min}}$$

D. Feature selection

In our proposed work, after the standardization, two methods of selection are used, namely RFE and murmur. MRMR feature selection includes features in terms of layouts and other features and class label compliance. RFE as the name suggests, also removes features and builds a model with the remaining features and tests the performance of the model. Selected features were trained in different algorithms that lead to increased efficiency of our proposed model.

E. Performance Evaluation Metrics for Model

After the feature is selected, the model is used and the output is generated in the form of an opportunity or category. The next step is to determine how well the model is using the test dataset based on specific metrics. In our work, testing the performance classification of different metrics such as accuracy, memory, precision, F-1 score, and AUC-ROC curve was used. Choosing the right metrics for testing a machine learning model is very important as it influences how performance is measured and compared.

F. Confusion Matrix

The confusion matrix is the most accurate matrix used to determine the accuracy and precision of the models. It is used for binary category and multi-stage problems. Describes the performance of class models where true values are already known. The confusion matrix is a table with two dimensions, one of the actual target value and one of the predicted values. To explain the concept of the matrix of confusion, consider the problem of binary division where classes 1 and 0 are shown in Figure 2.

The labels themselves are represented by rows and the predicted labels are indicated by columns. The basic conditions for confusion in the Matrix are discussed as follows:

True Positive (TP): Can be defined as the system's ability to properly classify conditions as positive, which means that if the actual label is 1, then the predicted label is also 1. As a percentage, it is expressed as a Positive Rate (TPR), also called empathy as part of a well-defined good example. It is given by:

$$\text{TPR (Sensitivity)} = \frac{TP}{TP+FN}$$

	Predicted label (1)	Predicted label (0)
Actual label (1)	True positive	False negative
Actual label (0)	False positive	True negative

Fig. 2. Confusion Matrix for classification

True Negative (TN): It is also called specification and is defined as the knowledge of the system to properly classify examples as negative, which means that if the actual label is 0 then the predicted label is also 0. It is represented as the True Negative Rate (TNR) in the terms of the segment of negative samples that are accurately predicted by the model.

False Positive (FP): In this case, the model unfairly classifies the conditions as straightforward. That is, the model predicts the class label as 1 whose name was 0. False Positive Rate (FPR) is represented as part of the worst cases predicted as positive and reported by:

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

False Negative (FN): It is the system's ability to incorrectly classify the examples as negative which means that for the actual label 1, the predicted label for a class is 0. False Negative Rate (FNR) is the fraction of positive samples

Engineering and Technology Journal for Research and Innovation (ETJRI)
ISSN 2581-8678, Volume IV, Issue I, Jan 2022

that were predicted as negative instances and is given by:

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

Precision: It is defined as the ratio of true positive relevant instances to the total number of retrieved instances. It is given by:

$$\text{Precision (P)} = \frac{TP}{TP+FP}$$

Recall: It is also called as the sensitivity and is defined as the fraction of correct positive examples predicted to the total number of positive occurrences.

$$\text{Recall (r)} = \frac{TP}{TP+FN}$$

F1-Score: Precision and recall are summarized into another metric which is called as F1- score. It represents a harmonic mean of recall and

$$F1-\text{score} = \frac{2-r-p}{r+p}$$

It may also be represented as:

$$F1-\text{score} = \frac{2-TP}{2-TP+FP+FN}$$

Accuracy: It is the fraction of the number of correct predicted examples to the total number of instances present in the dataset. It is given by:

$$\text{Accuracy} = \frac{\text{number of correct predictions}}{\text{Total number of predictions}}$$

For binary classification, it is expressed as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

IV. EXPLORNIG DATA ANALYSIS

I) Dataset:

Our dataset represents the 195 rows and 23 columns. Which shows that there are 195 persons are tested and various features which are used to detect the Parkinson's disease have been detected. The status bar graph diagram shows that numbers of patients are suffering from Parkinson's disease from the given dataset. It represents the overall true value of the model.

1-Represents the number of persons are suffering from the Parkinson's disease.

0-Represents the number of persons have healthy factor

iii) Heat Map

To determine the highly correlated features and low correlated features among the 23 features set the heat map has been visualized. BY using this heat map we can understand that spread 1, spread 2, PPE, HNR, MDVP (Flo), MDVP(Of).

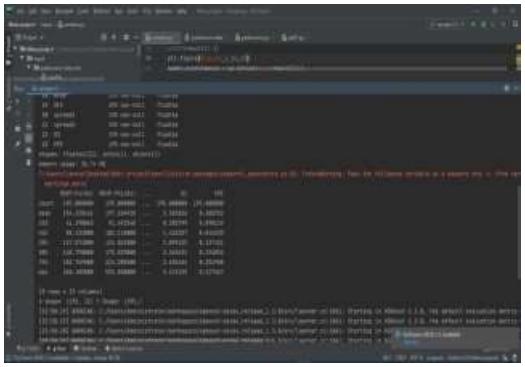

Fig. 3. Data set

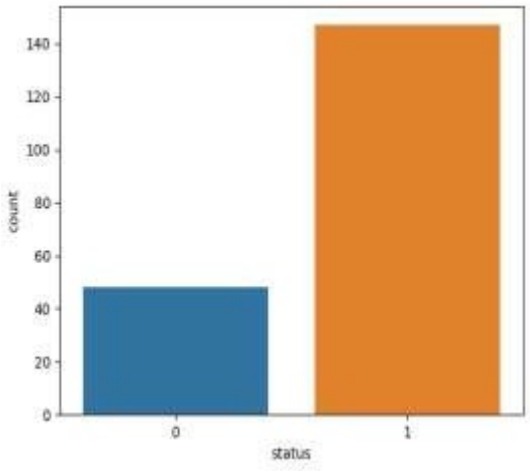

Fig. 4. Status

IV) Pair Plot

By using the heat-map we identify the correlated features which are highly contribute the detection of Parkinson's disease detection. Therefore pair plot has been drawn for the features to describe its detection.

Engineering and Technology Journal for Research and Innovation (ETJRI)
ISSN 2581-8678, Volume IV, Issue I, Jan 2022

V) Model Performance:

By using XG- Boost Algorithm XGB Classifier we trained and tested the dataset and found out its Accuracy and F1 Score rates with 93% and 95% respectively.

V. RESULT

To conclude our paper we compared our results with various machine learning Algorithm such as Logistic Regression model, Support vector machine model and Decision Tree Classifier model. By comparing the F1-Score of the above selected algorithm the comparison charts has been darkened. From this we can understand that XG Boost Algorithm have more accuracy prediction than any other algorithm models.

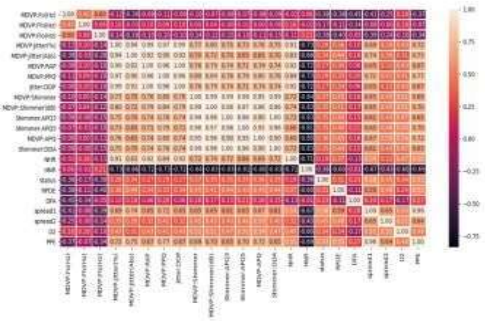

Fig. 5. Heat map

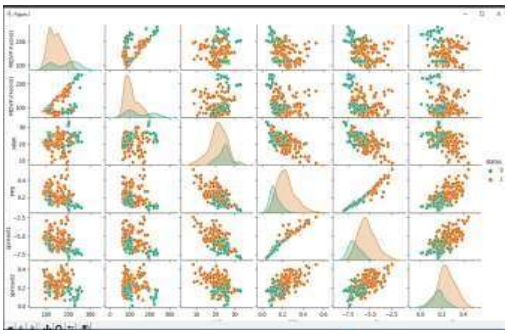

Fig. 6. Pair plot

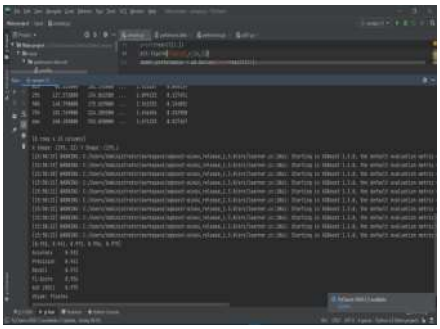

Fig. 7. Model performance

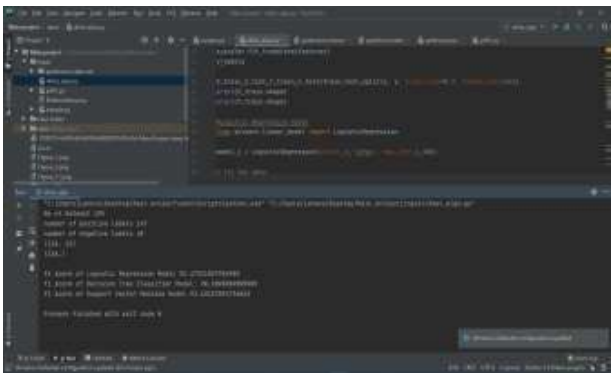

Fig. 8. Comparison results

VI. CONCLUSIONS

Currently, the Parkinson's disease research area is of much significance and its detection at the early stage can make the patient's life better.

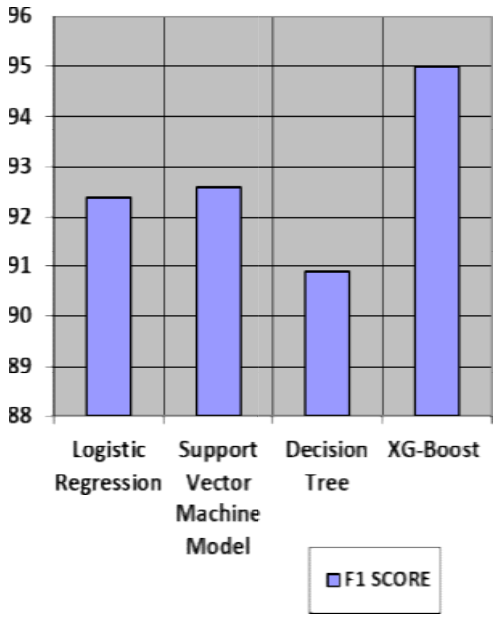

Fig. 9. Comparison charts

The recent developments in the methodologies through speech analysis have produced significant results. In our work, the problem of identification of Parkinson's disease is coped through a machine learning approach and different types of machine learning models have been employed for its detection. The main aim of this work is to show the PD diagnosis by analyzing the voice signals. From many years, speech processing has an incredible potential in the detection of PD as voice measurements are non-invasive. This work is intended to ascertain and analyses the performance of various classification algorithms. The different

classifiers were applied on a voice dataset and various evaluation metrics have been compared based on visualization and statistical analysis. Among all the classifiers, it was found that the Boost outperforms the other classifiers in machine learning algorithms. An accuracy of 92.76% was reported by using RFE feature selection technique while the accuracy of 95.39% was reported when using the murmur feature selection technique on all feature subsets which is higher than all state-of-art methods. Based on the results, the followings may be recommended:

(i) The Extreme Gradient Boost (Boost) technique should be used to develop model for PD detection problems.
(ii) As the features initially available to the system can be numerous, hence it is highly advisable to apply some feature reduction / selection technique to reduce to the complexity of the detection system.
(iii) The murmur technique assisted in achieving better results in our case, hence is strongly recommended for feature selection task.

Though the model works efficiently, this is limited by the richness hence a dataset with more no of samples would help the model generalize better. The proposed model is thus a reliable model to detect Parkinson's disease due to its efficient precision, F1-score, recall, and accuracy rates.

REFERENCE

[1] Manuel Gil-Martín, Juan Manuel Montero and Rubén San-Segundo (August 2019) Parkinson's disease Detection from Drawing Movements Using Convolutional Neural Networks (2019)
[2] Jorge Garza-Ulloa (MAY 2019) Update on Parkinson's disease, Am J Biomed Sci&Res - 2019 ISNS 2542-1747
[3] M. Beudel, P. Brown, Adaptive deep brain stimulation in Parkinson's disease (2015), 1353-8020/© 2015 The Authors. Published by Elsevier Ltd.
[4] Marla M. van Hymen a , Maria Florala Contagion a, b , Hub A.M. Middelkoop a, c ,Jacobus J. van Hilten a , Victor J. Geraedts a, d,Effect of deep brain stimulation on caregivers of patients with Parkinson's disease: A systematic review(2020),1353-8020/© 2020 The Authors. Published by Elsevier Ltd.
[5] A. Inguanzo , R. Sala-Llonch , B. Segura ,H. Erostarbe , A. Abos , A. Campabadal C. Uribe, H.C. Baggio , Y. Compta , M.J.Marti , F. Valldeoriola , N. Bargallo , C. Junque November(2020) 1353-8020/© 2020 The Authors. Published by Elsevier Ltd.
[6] Lisanne J. Dommershuijsen , Alis Heshmatollah M. Arfan Ikram a , M. Kamran
[7] Ikram (2020),Life expectancy of parkinsonism patients in the general population 1353-8020/ © 2020 The Authors. Published by Elsevier Ltd.
[8] Tarigoppula V.S Sriram1, M. Venkateswara Rao2, G V Satya Narayana3 , DSVGK Kaladhar4,"Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms", Volume 3 September 2013
[9] Karthikeyan.M 1, Chaitanya Rajeev Myakala 2,Sai Chaitanya Chappidi 3," Heart Attack
[10] Prediction Using XGBoost", Vol. 29, No. 6, 2020
[11] Nayan Kumar Sinha, Menuka Khulal, Manzil Gurung, Arvind Lal,"Developing A Web based System for Breast Cancer Prediction using XGboost Classifier", Vol. 9 Issue 06, June 2020
[12] Neharika D Bala, Anusuya S , "Machine Learning Algorithms for Detection of Parkinson's Disease using Motor Symptoms: Speech and Tremor", Volume-8 Issue-6, March 2020
[13] Timothy J. Wroge1 , Yasin Ozkanca ¨ 2 , Cenk Demiroglu2 , Dong Si3 , David C. Atkins4 and Reza Hosseini Ghomi4," Parkinson's Disease Diagnosis Using Machine Learning and Voice",November 2018