iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VII **Month of publication:** July 2020

DOI: http://doi.org/10.22214/ijraset.2020.7057

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VII July 2020- Available at www.ijraset.com

A Comparative Analysis for Prediction of Parkinson's Diseases using Classification Algorithm

Smriti Shikha[1], M. Arvindhan[2]
[1, 2]*Galgotias University*

Abstract: ***Machine learning has secured a significant name in health care sector because of its ability of improving the accuracy and time for disease prediction. Recently, Parkinson's is a noteworthy chronic diseases worldwide. It is observed that more than one million cases are common in India .There is the chances that 1.2 million people will be suffering with this diseases in the US at the end of 2030. Thus, early stage prediction of people's parkinson's severity is important in order to make fast planning of necessary treatment. In our study, we proposed a framework which will be helpful in real time prediction of parkinsons. We have use UCI Machine learning repository dataset which contains the acoustic features of voice recordings. Dataset is divided into 8:2 ratio as train and test data respectively. We use four important machine learning classfication technique i.e. SVM, Logistic regression, Extra tree classifier, Decision tree classifier for predicting parkinsons. 80% of the dataset is used to train the model for specified classification technique and validation is done over rest 20% of data. The performance measurement of the proposed classification technique was evaluated by applying the 10-Fold cross validation technique. The result shows that svm provides highest performance with tha accuracy 81%.***

Keywords: parkinson's diseases, logistic regression, svm, decision tree classifier, extra tree classifier, accuracy

I. INTRODUCTION

Parkinson's Diseases disrupt the systematic functioning in central nerve. It causes lack in proper functioning of body parts as well as internal cells. It mostly effects the movement of a person. It is progressive that means if it is not been cured in time then it will become worse day by day. Mostly, it effects the parts of brain where nerve cell become less functioning or may die. This reduces the efficiency of walking, talking, etc.It causes pain in joints, stiffness in the joints or limbs,or problems to the central part of the body mostly with spinal. These all mentioned symptoms doesn't gurantees that a person is surely having a parkinson's as this symptoms might be of other diseases too. The symptoms of parkinson's are of two type motor and non-motor. Motor symptoms are those which can be easily noticed by human. These include bradykinesia (slowness and shortness in movement), stiffness, rigidity, tremors at rest.

About more than 50,000 Americans are identified as having parkinson's. However, it is not necessary that this number is accurate because of the similarity of symptoms of parkinson's with other diseases. Also, many people takes it as a sign of normal aging and do not go under medical treatment. Due to its similarity of symptoms with other diseases sometimes person is incorrectly identify as having parkinson. While in the other hand person having parkinson is incorrectly identified as having other disease.

It has been found that its occurrence is higher in developed countries. In rural areas , people living in areas with pesticide use are also prone to it. Parkinson's occurs when nerve system dies or impaired which results in slow functioning. Many brains parts are effected due to this impairment. The most symptoms are because due to loss of neurons in substantia -nigra. A chemical called 'Dopamine' is produced by neuron in brain. It is useful in transmitting a message between substantia nigra and corpus stratium which helps in efficient movement. Due to loss in dopamine, a irregular nerve firing patterns is generated in the brain which cause improper movement.

Staging of parkinson's diseases are as follows:

1) At stage one, symptoms can be seen in either side of the body
2) At stage two, symptoms start to hit on both side of body
3) At stage three, person start to feel difficulty in balancing body.
4) At stage four, disabilty starts to hit the body. although a person is able to work or stand improper.
5) At stage five, person are require to use wheel-chair or ask for complete bed rest.

This article is organize in following way. Section 2. describes the materials and methodology. Section 3 illustrate the result and discussion of this work. Section 4 describes conclusion from this entire work.

II. MATERIALS AND METHODOLOGY

In this section, Section 2.1 contains the description of datasets. Section 2.2 describes the proposed framework for parkinson's diseases. Section 2.3 describes the classification algorithms used in this study.

A. Datesets

The Replicated Acoustic Features-Parkinson Database Dataset is used in this proposed work, available in the UCI Machine Learning Repository. The standard speech tests records were conducted and from there speech sounds were recorded and analysed using praat software [9].The dataset consists of some vocal fundamental frequency, recording, amplitude variations, status, nonlinear fundamental frequency measures.

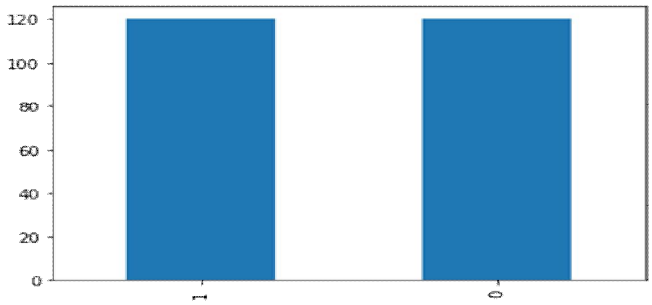


Fig.1. Bar plot for column name status.

This dataset has 240 persons voice data with 46 columns and is 120 KB. . All the columns are of type float64 except status, gender, recordings, these are of type int64 and id is of type object . The bar plot in Fig 2 shows the relationship between a numeric and a categoric variable for the column status.

Description of dataset:

1) Id – Subject's Identifier
2) Recording – No of recording
3) Status – categorical value 0 and 1. 0 for persons who are not having a parkinson's and 1 for those suffering from parkinson's.
4) Gender – 1 shows recording is of female and 0 shows recording is of male.
5) Pitch local perturbation measures – Jitter_rel specifies relative jitter, Jitter_abs specifies absolute jitter, Jitter RAP specifies relative average perturbation and Jitter_PPQ specifies pitch perturbation quotient.
6) Amplitude perturbation measures – Shim_loc specifies local shimmer, Shim_dB specifies shimmer in dB, Shim_APQ3 specifies 3-point amplitude perturbation quotient,Shim_APQ5 specifies 5-point amplitude perturbation quotient, Shim_APQ11 specifies 11-point amplitude perturbation quotient.
7) Ratio of Harmonic with respect to noise.
8) Mel frequency cepstral coefficient-based spectral measures of order 0 to 12
9) RPDE- It is a measure of dynamical complexity
10) DFA – Signal Factor scaling exponent
11) PPE - Three nonlinear fundamental frequency variation measures
12) GNE – Ratio of Glottal to noise excitation..

B. Proposed Framework for Parkinson's Diseases.

In this segment, the framework of proposed work is presented. The proposed framework consists of six steps

1) Getting available with data.
2) Data Preprocessing.
3) Data Exploration.
4) Applying classification algorithm
5) Performance evaluation on various measures
6) Comparision on the basis of accuracy
7) Results.

a) A framework for Parkinson's prediction using classification technique is proposed. The proposed work will help doctors in an correct decision making regarding possibilities of this diseases.

b) The parkinson's diseases is growing enormously, the proposed model will try to break the increase in severity of that disease by early prediction.

c) As we have seen that maximum study or papers do not involve F-score, recall and precision etc. But, in our study, we have included the confusion matrix and precision, recall, F-score etc.

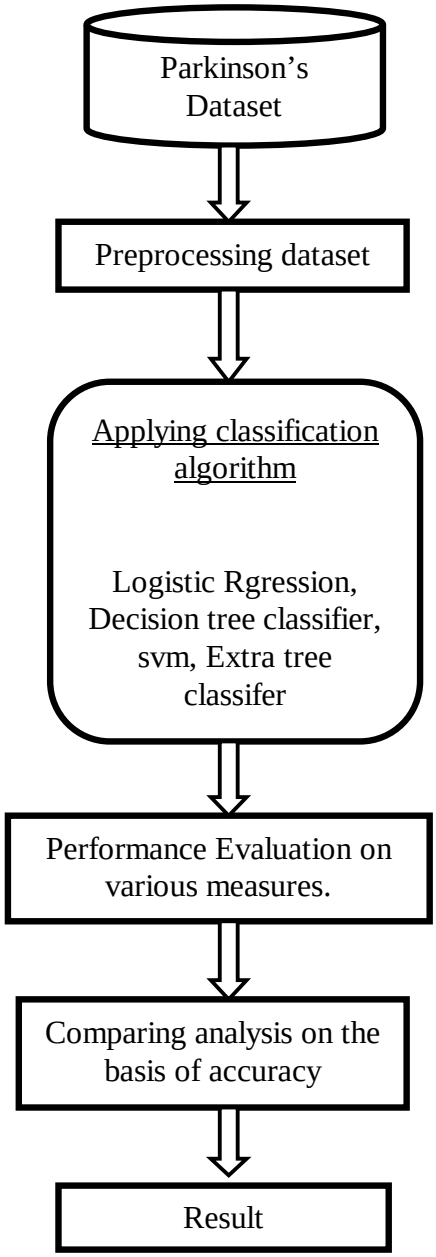


Fig.2. Proposed framework for parkinson's diseases.

We start with importing the necessary libraries required for the proposed work.Those libraries are Pandas, Matplotlib, Seaborn, Numpy, Standard Scaler, Train_test_split, Logistic Regression, Extra Tree Classifier, SVC, Decision Tree Classifier. Accuracy_score. After this dataset i.e. ReplicatedAcousticFeatures-ParkinsonDatabase.csv is read in working environment. Next comes the most important part of the work that is Data preprocessing. It is necessary for a data to be cleaned as noisy data may affects adversly in creating model for prediction as it may decrease the quality of model and hence decreasing the result performance.

Following are the steps used in preprocessing the parkinson's dataset for getting correct result.

i) Finding null values. In our work, it gives false which means our dataset has no missing value. If it gives True, we would have go by the step of dropping values

ii) Assigning data as label and features

iii) Detecting outliers. It is seen by plotting the bar plot of status it shows no outliers

iv) Scaling the dataset.

v) Next step is splitting the dataset. The Dataset is divided into features (i.e. all the columns except status) and labels (i.e. column Status) The imported dataset is splitted into train and test data. in 8:2 ratio through the help of train_test_split. 8:2 is training data with respect to testing data. Check the outliers through boxplot.

vi) Final step is the prediction using the machine learning algorithm i.e. logistic regression, svm, extra tree classifier, decision tree classifier. The accuracy are computed. The accuracy of all the implemented algorithm are compared and result is concluded which algorithm has the highest accuracy. That algorithm can be used in real time predictions.

C. Description of classification algorithm used in the proposed model

1) *SVM:* It stands for support vector machine. It is type of learning which belongs to supervised class and used in classification. It also can be used in regression problem. It can be applied to both linear and non-linear problems. It is useful as it works good for many health care data. The algorithm starts with plotting the each value of data in n dimensional space. Here, n is the count of the features in the data. Then, performs the classification where we find the most suitable hyper-planes which classify the classes greatly. It is widely used because of its accuracy. It works well with the smaller size of datasets.

2) *Decision Tree Classifier*: Decision tree is a supervised machine learning algorithm. In this, data is continuosly partitioned according to certain parameter. Decision Tree consists of Nodes, Edges or branch, Leaf nodes. Nodes test for the value of a certain attribute. Edges are the connection between two nodes. Leaf nodes represents the outcome. Decision tree are of mainly two types classification trees and Regression trees. Classification tree are those where the outcome is categorical/discrete. This type of tree involves building process called binary recursive partitioning. Regression tree are those where decision variable are continous values. Decision tree classification have some advantages over other classification techniques like easy to interpret, doesn't take useless features into model building and are extremely fast.

3) *Extra Tree Classifier:* It is an ensemble learning technique. It is similar to random forest but diviates to random forest in terms of constructing decision tree in the forest. The classification result is an aggregate of the multiple de-correlated decision trees collected in forest.

4) *Logistic Regression:* It is supervised learning algorithm. It helps to find the probability of dependent variable.the dependent variable must be dichotomous that is it can be only two classes. In our study,dependent variable is status which has class either 0 or 1.

III. RESULT AND DISCUSSION

A. Measurements of Classification Algorithm

In this study, for measuring the performance of classification algorithm, we have use 10-fold validation technique. It provide us the statistical measurements which helps in evaluating algorithm performance. Stastical measures are accuracy, precision, sensitivity, specificity. These classification are calculated by the following terms. TP, FP, TN, FN.

TP stand for True positive where predicted answer is yes and a person have also parkinson's diseases.

FP stand for False positive where predicted answer is yes and a person is not suffering from parkinson's diseases.

TN stands for True positive where predicted answer is no and a person is suffering from parkinson's diseases.

FN stands for False positive where predicted answer is no and a person have parkinson's diseases.

Accuracy is ratio of summation of TP and TN to the summation of TP, TN, FP, FN.

Sensitivity is also called recall and TPR. It is calculated by taking the ratio of TP to the summation of TP and FN.

Specifity is also called TNR .It is suitable oppposite of recall which is ratio of TN to the summation of TN and FP.

Precision is the ratio of TP to the summation of TP and FP.

F1 is calculated by taking ratio of double of recall multiplied by Precision to the summation of recall and precision.

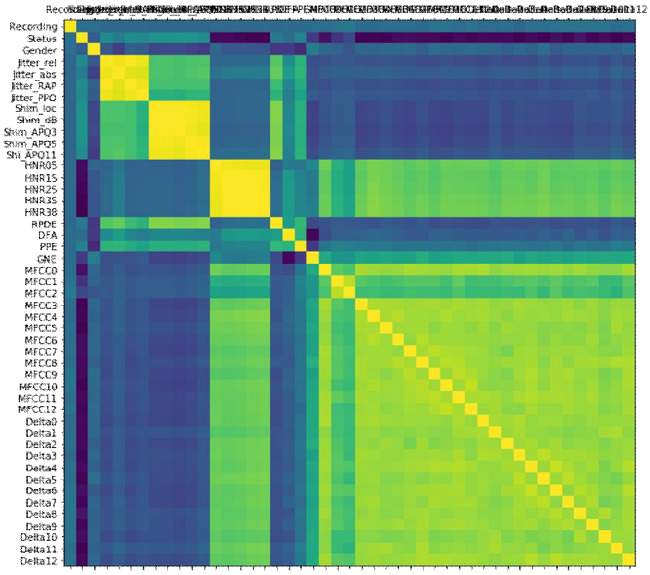


Fig.3. Heat map for checking coorelated columns.

B. Analysis of results

In our study, we have done analysis to compare the performance of the used machine leaning algorithms for parkinson's prediction. From the parkinson's dataset, 120 were positive and 120 were negative .We split the dataset in two parts where 80% is training and testing part contain 20%. The barplot of label is plotted in order to check

for the outliers shown in Fig. 1. From the boxplot plotted for each variable evident that if a patient has a lower rate of 'HNR' then they are affected by parkinson's Diseases. To drop the coorelated columns dataset was checked for corelated columns. The heatmap are shown in Fig. 3 for all columns.

Fig. 4. depicts the performance of used classification algorithms for parkinson's prediction. Here, SVM performance well in terms of highest accuracy of 81 %.

However, Decision tree classification performance was lowest in comparision to others.

Table 1 elaborates the performance of classification algorithm implemented in our work. It illustrates the different classification measures.

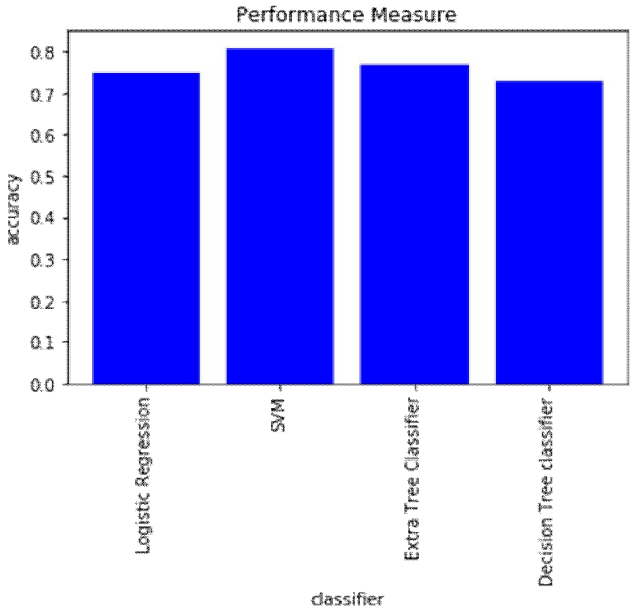


Fig .4. Performance measure of classifiers

Table -1: classification performance measures

Measurement Techniques	LR	SVM	ETC	DTC
Accuracy	0.75	0.81	0.77	0.73
Precision	0.75	0.81	0.77	0.73
Sensitivity	0.75	0.81	0.77	0.73
F-1	0.75	0.81	0.77	0.73

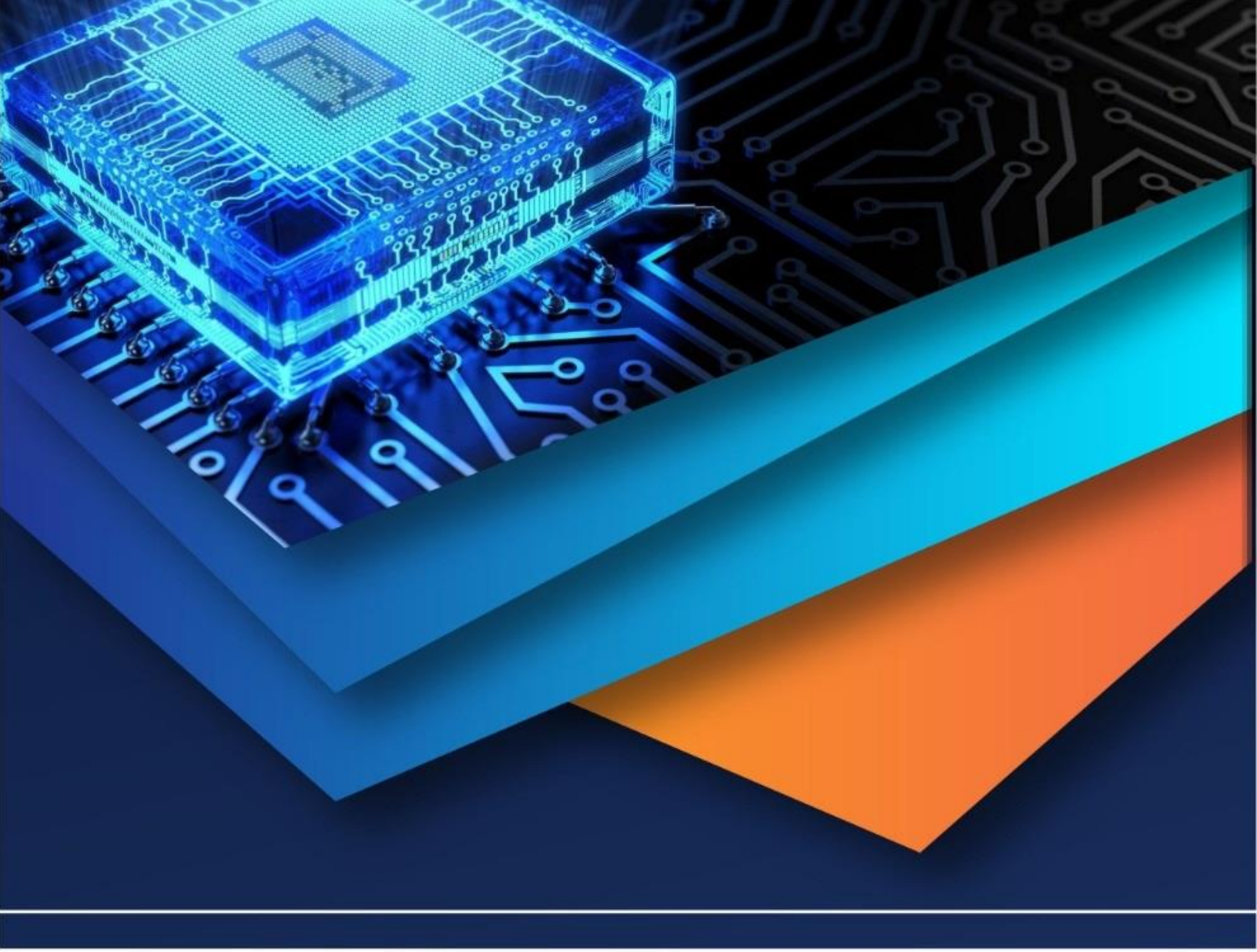
In Logistic regression , the accuracy is about 0.75. In SVM, accuracy_score comes out to be 0.81. In decision tree classifier, the accuracy is about 0.73 and for Extra tree classifier it is about 0.77. It can viwed that all algorithms performed well while svm gives the best performance with 81 % accuracy. Hence, this model can be used to identify the possibility of having parkinson's diseases with high level of surity of correct prediction.

IV. CONCLUSION

The main contribution of this study are as follows, firstly, we have compare the performance of the machine learning algorithm used and evaluate those performance using 10-fold validation technique. Secondly we propose a framework for prediction of parkinson's diseases.The result shows that among all the experimented algorithm svm exhibit the highest performance with accuracy of 81%.Since,the prediction of this diseases is not direct which means we can't predict it using any test. So, this study provides the doctor with high level of accuracy in decision making and clinical diagnosis. The proposed model can be used in real life application.

REFERENCES

[1] Patra, A. K., Ray, R., Abdu, A. A., & Dash, S. R. (2019). Prediction of Parkinson's disease using Ensemble Machine Learning classification from acoustic analysis. *International Conference on Biomedical Engineering*. Malaysia.
[2] Challa, k. N. (2016). An improved Approach for prediction of Parkinson's Diseases using Machine Learning Techniques. *International conference on Signal Processing, Communication, Power and Embedded System*.
[3] Kokilam , V., Latha, & Latha, D. P. (2016). A comparative study of parkinson's disease prediction using classification and filtering techniques. *International Journal of Applied Engineering Research*.
[4] Rastegar, D. A., Ho, N., Halliday , G. M., & Dzamko, N. (2019). Parkinson's progression prediction using machine learning and serum cytokines.
[5] Patra, A. K., Ray, R., Abdu, A. A., & Dash, S. R. (2019). Prediction of Parkinson's disease using Ensemble Machine Learning classification from acoustic analysis. *International Conference on Biomedical Engineering*. Malaysia.
[6] Smith, B. (n.d.). An approach to graphs of linear forms (Unpublished work style). unpublished.
[7] S, A., & Vincent, K. (2019). Identification of Parkinson Disease Patients Classification using Feed Forward Technique Based On Speech Signals. *International Journal of Engineering and Advanced Technology (IJEAT)* .
[8] Schrag, A., Anastasiou, Z., Ambler, G., Noyce, A., & Kate Walters, K. (2019). Predicting diagnosis of Parkinsons disease:A risk algorithm based on primary care presentations.
[9] Kannan, A. R., Kiruthiga, M., Kumar, C. P., Dhenaskumar, A., & Manoj, S. (2019). Prediction of Parkinson's diseases using Machine learning Techniques.
[10] Schrag, A., Anastasiou, Z., Ambler, G., Noyce, A., & Kate Walters, K. (2019). Predicting diagnosis of Parkinsons disease:A risk algorithm based on primary care presentations.
[11] Little, M. A., McSharry, P. E., Roberts, S. J., Costello , D. A., & Moroz, I. M. (2007). Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection.
[12] M.Arvindhan,A.Anand,"Scheming an Proficient Auto Scaling Technique for Minimizing Response Time in Load Balancing on Amazon AWS Cloud"Available at SSRN 3390801,2019
[13] Clustering algorithm networks test cost sensitivity for specialist divisions, DOI: 10.21917/ijivp.2019.0298 VOLUME: 10,ISSUE: 02
[14] Aravindhan, M., & Ande, B. P. (2020). Data mining approach and security over Ddos attacks. *ICTACT jouranl on soft computing*.
[15] Wrogel, T. J., Ozkanca, Y., Demiroglu, C., Si, D., Atkins, D. C., &(2019). Ghomi, R. H. (2018). Parkinson's Disease Diagnosis Using Machine Learning and Voice Timothy. *IEEE*.
[16] Liaw, A., & M. Wiener et al. (2002). Classification and regression by .randomforest. 18-22
[17] Pedregosa, F., Varoquax, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, D. (2011).
[18] Mandal, I., & Sairam, N. (2012). New machine- learning algorithms for prediction of Parkinsons disease. *International journal of systems* on *science*.
[19] Machine Learning in Python. Journal of Machine Learning Research 12 (2011) 2825-2830, 2825-2830.
[20] Little, M. A., McSharry, P. E., Hunter, E. J., Ramig, L. O., & Spielman , J. (2008). Suitability of dysphonia measurements for telemonitoring of Parkinson's.
[21] R, P., SD, R., Pk, M., & S., G. (2014). Parkinson's disease detection using olfactory loss and REM sleep disorder features. *IEEE Eng Med Biol Soc*.
[22] Rastegar, D. A., Ho, N., Halliday , G. M., & Dzamko, N. (2019). Parkinson's progression prediction using machine learning and serum cytokines.
[23] github. (n.d). Retrived from https://github.com/razuswe/Prima-Indian-Diabetes-prediction/blob/master/performance.ipynb.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY