ISSN 2347 - 3983
Volume 8. No. 4, April 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter31842020.pdf
https://doi.org/10.30534/ijeter/2020/31842020

Detecting Central Nervous System Disorder Using Machine Learning Technique (XGB Classifier)

Sri Lasya Dharmapuri[1], Pavan Kumar Dandamudi[2], Vinoothna Manohar Botcha[3], Bhanu Prakash Kolla[4]
[1,2,3]Koneru Lakshmaiah Education Foundation, India.
lasyadharmapuri@gmail.com[1], pavan.dandamudi.PK@gmail.com[2] , manoharvinoothna@gmail.com[3],
drkbp@kluniversity.in[4]

ABSTRACT

Parkinson's disease is a nervous system disorder which is progressive effects the aged humans (above 50) brain movements. Parkinson's disease (PD) a neurodegenerative disorder affects nearly a million numbers of people in the US by "PARKINSON'S FOUNDATION Understanding Parkinson". The body condition of the affected person is diagnosed and the symptoms of difficulty in pronouncing, pulse rate, pulse amplitude are some of the features taken into consideration for the detection of Parkinson using a machine learning technique.in this, supervised machine learning known as EXTREME GRADIENT BOOST Classifier is used for the prediction of disease based on the symptoms.

Key words : Parkinson disease, Machine Learning, Extreme gradient Boosting.

1. INTRODUCTION

Parkinson is the central nervous issue, therapeutically portrayed as a neurological disorder by James Parkinson in the year 1817[1]. It is the second regular neurological ailment in which a territory of the cerebrum is harmed. The specific reason for Parkinson's malady isn't known. It may cause because of both hereditary and natural effects. It causes different signs and manifestations. These manifestations are declining after some time. Numerous specialists utilize the Hoehn and Yahr scale to arrange their stages [2]. The disease is regularly found between the age of 50-60 years. These signs and indications are assembled into Motor Symptoms and Non-Motor manifestations. Engine side effects are those which influence development and muscles. Non-engine is neurobehavioral psychological issues, rest issues, tangible issues and autonomic neuropathy (dysautonomia)[3].

Speech disturbances are the normal motor symptom in Parkinson [3] and about 90% of individuals are influenced by speech impedance [4].signal processing on their voice is minimal effort and suitable in telemonitoring and telediagnosis frameworks. Speech disturbances are principally hypophonia-relax speech because of shortcoming in vocal musculature, monotonic speech-manages speech quality(soft, dry and dreary) and festination speech-when speech turns out to be too much quick, delicate, hoarse and poor intelligible[3]. In running speech few states standard sentence incorporates agent semantic units' gives debilitation indications in the vocal issue.[5]

2. RELATED WORK

Detection of Parkinson's ailment using the discourse unsettling influences of vocal issue using a machine learning strategy by considering acoustic estimations of dysphonia as successful highlights [7-14]. This field can be arranged into (1) endeavors to discover best vocal highlights and produce new datasets, (2) those attempts to discover increasingly powerful highlights from the existing dataset and attempt to improve exactness [6, 15]. In [13], Sakar et al. introduced a dataset of 40 subjects including 20 PD. Every individual is prepared to state a lot of 26 unmistakable vowels, words, numbers and short sentences. They applied s-LOO approval strategy known as Summarized leave-one-outs in which all voice tests of each individual condensed with scattering measurements of mean, middle, go, standard deviation, cut mean, interquartile range and mean total blunder, their exactness got as 77.5%. Tsanas et al.[10] concentrated on checking PD Progression of separated highlights from an immense dataset of 6000 voice tests that are applied for machine learning procedures, from 42 patients with beginning period PD endeavoured to assess bound together Parkinson's sickness rating scale(UPDRS) for both straight and non-direct relapse demonstrating an exactness of 7.5 focuses significantly more than from clinical UPDRS estimation. By utilizing fake neural systems; it is difficult to get exact consequences of above 90% [9].there are three datasets accessible for PD discourse-based regions. Sakar and Kursun [15] attempted to survey the pertinence and relationship between's the highlights and PD Score exposed to common data-based choice calculation with stage test and feed the information with chose highlights positioned on greatest importance least repetition (mRMR) into a Support vector classifier utilizing Leave-one-subject-out(LOSO) as the cross-approval procedure to maintain a strategic distance from predisposition.

Their approach increasing accuracy of 92.75% [8]. Shahbaba and Neal [16] speaks to a non-direct model dependent on Dirichlet blends and acquired a classification accuracy of 87.7%. Right now non-direct change combined with the SVM approach having the best accuracy of 93.47%.PD diagnosis has been performed utilizing deep neural network classifier with 30 diverse 10-overlap cross-approval having stack encoder and a softmax picking up accuracy 93.79(Caliskan, A.,2017).[17-20]

3. THEORETICAL ANALYSIS

3.1. Materials (Data)

The downloaded dataset having subjects 5876 of PD Patients with 24 features. Each theme, 20 diverse sound recordings have assembled the highlights accumulated and a portion of the highlights can be portrayed as follows.

Table 1: Description of the dataset

Feature Number	Feature Name	Feature description
1	age	Age of the subject
2	Test_time	Time since recruitment in the trail.
3	Motor_UPDRS	-
4	Total_UPDRS	-
5	Jitter(Absolute)	Frequency parameters
6	Jitter: RAP	
7	Jitter:PPQ5	
8	Jitter: DDP	
9	Shimmer	Amplitude parameters
10	Shimmer(dB)	
11	Shimmer:APQ3	
12	Shimmer:APQ5	
13	Shimmer:APQ11	
14	Shimmer:DDA	
15	NHR	Noise-to-Harmonic
16	HNR	Harmonic-to-Noise
17	RPDE	Recurrence period Density Entropy
18	DFA	Detrended Fluctuation Analysis
19	PPE	Pitch period entropy
20	Jitter(Percentage)	Frequency percentage

The data set composes of 24 features based on Baseline features. Status –The health status of the subject. For detecting the person is healthy or affected by the disease.

The basic early signs of the Parkinson's patient can be evaluated from the "Parkinson's foundation" labelled as (Parkinson's Foundation).

Figure 1: Early signs of Parkinson patient

Among all, the soft and low voice feature of a Parkinson patient can be easily estimated based on distinct letters to pronounce and analysed the pronunciation.[21]

3.2 Methods

The flow of data in the classification model is to collect the voice records that are gathered having both Parkinson's disease data and non-Parkinson's datasets.

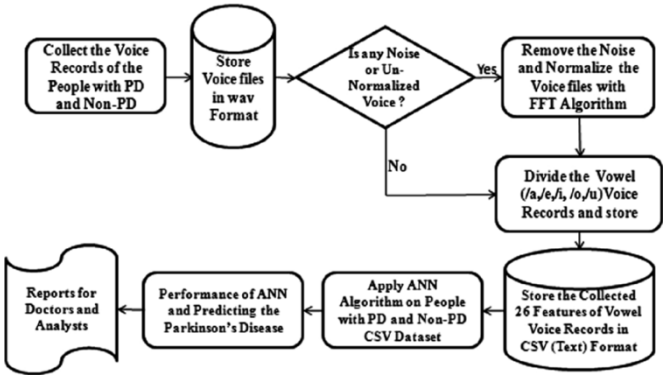

Figure 2: Flow diagram of the model

The features in Parkinson's disease dataset are taken for the analysis and can be estimated in histogram representation for the features A histogram represents the shape and spread of continuous data fall in a specific range.[22]

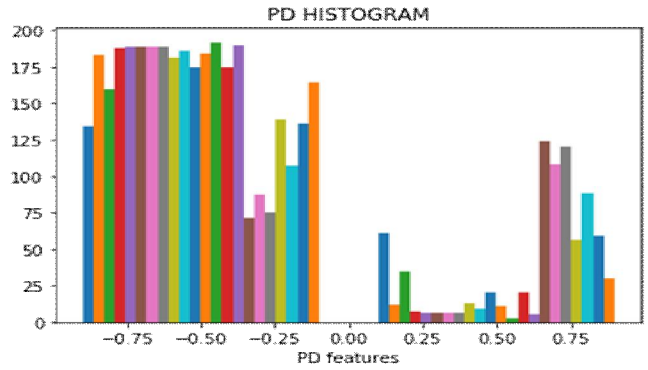

Figure 3: Histogram of Features

The total mean of the features is **-0.45009599942297557**. Each colour represents the density of features.

3.3 Data Pre-processing

This is an Essential cleaning and data representation step in knowledge discovery. The dataset having missing values or the voice signals with noise involvement are cleaned and pre-processing the features by features selection algorithms. a StandardScalar/MinMaxScaler is used in pre-processing step for the extraction of features. The redundant and incomplete information is to be validated and filtered for better performance in the model classification and prediction.

3.4 Feature Selection

This is the progression in pattern recognition to construct helpful data and a measure of relevance/dependence utilized in filter strategies for estimating relevance measures with the objective variable. The training clinical dataset needs to be prep-processed for feature selection, classification, prediction and analysis. The target classes are Motor_UPDRS and Total_UPDRS. Minimum Redundancy Maximum Relevance (mRMR) extracts the useful and relevant features in data with output by minimizing redundancy.

3.5 Model

A series of decision tree algorithms and various implementations were utilized to arrange dataset included standard random forest, GB trees and tree classifiers. The decision tree works with binary classification to isolate homogeneously by suing measurements that limit data entropy.

Extreme Gradient Boosting (XGB) Classifier Model

Extreme Gradient Boosting is one of the Ensemble machine learning technique of highly versatile and flexible that can work through classification, regression and ranking problems. A decision-tree technique that used a GB technique for prediction. For unstructured data formats like images and text, the artificial neural network is performed where for structured and tabular data, decision tree-based model is implemented for better results.

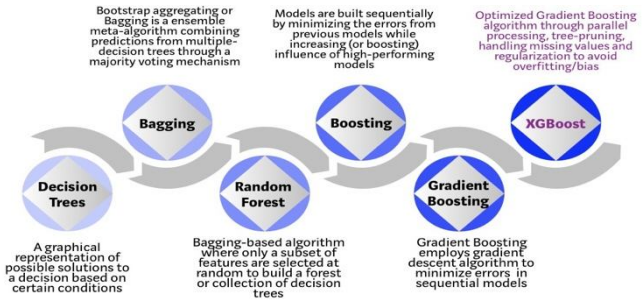

Figure 4: Classification Models

Many applications are used to perform regression, classification, ranking and required prediction analysis, for scalability, language scalability and cloud integration(AWS, Azure, Yarn clusters). It performs well because the principle of gradient boosting algorithm improves based on GBM framework through system optimization and algorithmic enhancement.

Key parameters in XGBoost

n-estimators = number of runs XGBoost,
learning rate = learning speed,
early_stopping_rounds = over fitting prevention,
stop early when improvement[12].

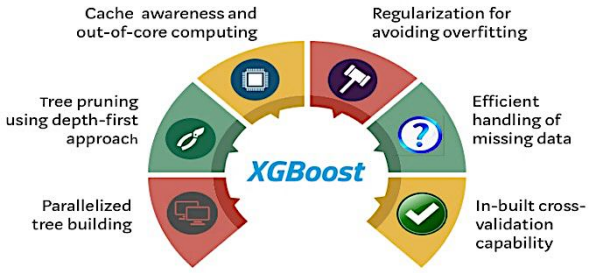

Figure 5: XGBoost Model

System optimization methods might be a Parallelization, tree pruning and equipment optimization and for Algorithmic Enhancement systems, Regularization, Sparsity Awareness weighted Quantile sketch and cross-approval.

Algorithmic Enhancements

1. **Regularization:** penalize the complex model by LASSO (L1) Regularization and Ridge (L2) regularization to avoid overfitting.
2. **Sparsity Awareness:** As XGBoost have sparse features as inputs by "learning" best missing values on training loss and maintains different sparse patterns.

3. **Cross-Validation:** cross-validation is implemented at each step to specify exact boosting iterations required for a single run [12].

When algorithms like logistic regression(LR), random forest(RF), standard GB and XGBoost performed on dataset, The XGBoost algorithm model performs well in accuracy metrics and processing time.

3.6 Evaluation metrics

Evaluating a model gives a satisfying result in the performance.in many cases; accuracy metric is used to measure the model performance.

Confusion Matrix

Table 2: Confusion Matrix

		Truth: Positive	Truth: Negative	
Test	Positive	True Positive	False Positive Type I α	Total Testing Positive
	Negative	False Negative Type II β	True Negative	Total Testing Negative
		Total Truly Positive	Total Truly Negative	Total

Our model is based on Binary classification with a prediction of resultant as "YES" or "NO", can be labelled as '1' or '0'.For this, the confusion matrix well performed in prediction.

Where

True Positive(TP) is the value, where were Patients correctly classified according to the model. *True Negative(TN)* is the value, where patients were healthy according to the model. *False Negative(FN)* is the value, where patients were wrongly diagnosed as healthy according to the model. *False Positive(FP)* is the value of healthy patients labeled as having PD by the classifier.

Based on the Positives and Negatives

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

The recall, precision and F-1 measured on training set for model selection.

Recall = $\frac{\text{True Positive}}{\text{True positive + False Negative}}$

Precision= $\frac{\text{True Positive}}{\text{True Positive + False Positive}}$

F-1: $\frac{2(\text{Recall} * \text{Precision})}{\text{Recall + Precision}}$ [22]

Classification Accuracy: Accuracy works well when sample of classes are equal.

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

of a classifier on the dataset is the percentage of test data that are being classified accurately and mathematically represented as

$$Accuracy : \frac{TP+TN}{TP+TN+FP+FN}$$

A known metric in machine learning which is reliable where TP, TN, FP, FN are taken into consideration. It is the correlation coefficient between observed (Actual) labels to predicted values by binary classifier.

$$\textbf{MCC} = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Gives a value ranges from -1 to 1

+1 indicate perfect prediction
0 indicates that the classifier is no better than randomly predicted.
-1 indicates disagreement (not in related) between actual and predicted.

4. RESULTS AND DISCUSSIONS

Initially, the dataset shape, information, and description of features are referred. Training and testing sets are considered from the features. The training dataset is trained for the prediction and testing data is for sample evaluating metric measures. After Training and Testing are performed; the Extreme Gradient Boosting algorithm classifier which performs ensemble technique (Boosting) with Gradient Search which is a heuristic algorithm for optimum (Best) results. The dataset has a "Status" column, having both Parkinson and non-Parkinson subjects. This classifier model is used to detect the subject having Parkinson which is set to be a Target.

Result in python implementation

XGB Classifier: **0.9591836734693877**

The classification accuracy approach is 95.91836734693877 approximately equals 96% of prediction about Parkinson's disease accurately.

The metric report can be obtained as

Precision	Recall	F-1	Measures	Support
0	1.00	0.80	0.89	10
1	0.95	1.00	0.97	39

5. CONCLUSION

Patients with Parkinson's (PWP) on their vocal samples detection have attractive research. Discriminating the Parkinson's patient from the healthy person's data based on the gathered samples, extracting information or creating a new dataset for the analysis which are used to classify the disease. By using a single classifier, the vocal terms can be classified with poor accuracy .the prosed test takes each feature and majority voting is considered to resolve the PD status for detecting. The features in discriminating are not in the same proportion. The dataset will be well designed based on their mental status and symptom conditions for having the best results in prediction. In the initial stages of this disease is better focused on prevention.

REFERENCES

1. Parkinson J 1817. An essay on the shaking palsy. Whittingham and Rowland for Sherwood, Needly and Jones, London [Google Scholar]
2. J. Jankovic, "Parkinson's disease: clinical features and diagnosis,"*Journal of Neurology, Neurosurgery and Psychiatry*, vol. 79,no. 4,pp. 368–376, 2007. https://doi.org/10.1136/jnnp.2007.131045
3. A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with Parkinson's disease," *Behavioural Neurology*, vol. 11, no. 3, pp.131–137, 1998 https://doi.org/10.1155/1999/327643
4. P.H. Dejonckere, P. Bradley, P. Clemente et al., "A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques: guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS)," *European Archives of Oto-Rhino-Laryngology*, vol. 258, no. 2, pp. 77–82, 2001. https://doi.org/10.1007/s004050000299
5. A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high accuracy classification of Parkinsons disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1264–1271,2012. https://doi.org/10.1109/TBME.2012.2183367
6. M. A. Little,P.E.McSharry, S. J. Roberts, D. A. E. Costello, and I. M.Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *BioMedical Engineering OnLine*, vol. 6, article 23, 2007. https://doi.org/10.1186/1475-925X-6-23
7. M. A. Little, P. E.McSharry, E. J. Hunter, J. Spielman, and L.O.Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1015–1022, 2009. https://doi.org/10.1109/TBME.2008.2005954
8. W.-L. Zuo, Z.-Y. Wang, T. Liu, and H.-L. Chen, "Effective detection of Parkinson's disease using an adaptive fuzzy knearest neighbor approach," *Biomedical Signal Processing and Control*, vol. 8, no. 4, pp. 364–373, 2013. https://doi.org/10.1016/j.bspc.2013.02.006
9. Empirical Study and Statistical Performance Analysis with ANN for Parkinson's Vowelized Data Set T. PanduRanga Vital Gorti Satyanarayana Murty K. Yogiswara Rao T. V. S. Sriram
10. Athanasios Tsanas*, Max A. Little, *Member, IEEE*, Patrick E. McSharry, *Senior Member, IEEE*, Lorraine O. Ramig, Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests, TBME-00652-2009.R1
11. *UCI Machine Learning Repository- Centre for Machine Learning and Intelligent System. http://mlr.cs.umass.edu/ml/datasets/Parkinsons+Telemonitoring*
12. K.B., Prakash. "Information extraction in current Indian web documents." International Journal of and Technology(UAE), 2018: 68-71. https://doi.org/10.14419/ijet.v7i2.8.10332
13. Prakash K.B., Dorai Rangaswamy M.A. "Content extraction studies using neural network and attribute generation." Indian Journal of Science and Technology, 2016: 1-10.
14. J. Jankovic, "Parkinson's disease: clinical features and diagnosis," *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 79, no. 4, pp. 368–376, 2007. https://doi.org/10.1136/jnnp.2007.131045
15. L. M. de Lau and M. M. Breteler, "Epidemiology of Parkinson's disease," *The Lancet Neurology*, vol. 5, no. 6, pp. 525–535, 2006. https://doi.org/10.1016/S1474-4422(06)70471-9
16. Ismail, M., Prakash, K.B. & Rao, M.N. 2018, "Collaborative filtering-based recommendation of online social voting", International Journal of Engineering and Technology(UAE), vol. 7, no. 3, pp. 1504-1507. https://doi.org/10.14419/ijet.v7i3.11630
17. K.B., Prakash. "Information extraction in current Indian web documents." International Journal of and Technology(UAE), 2018: 68-71. https://doi.org/10.14419/ijet.v7i2.8.10332
18. Prakash K.B., Dorai Rangaswamy M.A. "Content extraction studies using neural network and attribute

generation." Indian Journal of Science and Technology, 2016: 1-10.

19. Prakash K.B., Dorai Rangaswamy M.A., Raman A.R. "Text studies towards multi-lingual content mining for web communication." Proceedings of the 2nd International Conference on Trendz in Information Sciences and Computing, TISC-2010, 2010: 28-31. https://doi.org/10.1109/TISC.2010.5714601
20. Prakash K.B., Rangaswamy M.A.D. "Content extraction of biological datasets using soft computing techniques.", Journal of Medical Imaging and Health Informatics, 2016: 932-936.
21. Botcha, V.M., Monitha, G., Madala, D.N.S., Kolla, B.P. "ANALYSIS OF NATURE INSPIRED ALGORITHMS". Journal of Critical Reviews. Vol 7, Issue 4, 2020. https://doi.org/10.31838/jcr.07.04.140
22. Botcha, V.M., Koll, B.P. "Predicting breast cancer using modern data science methodology". International Journal of Innovative Technology and Exploring Engineering(IJITEE),ISSN: 2278-3075, Volume-8 Issue-10, August 2019. https://doi.org/10.35940/ijitee.J1077.0881019