TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 4288-4293

An Early Prediction of Parkinson's Disease using Machine Learning Techniques

[1]Basavaraj S. Hadimani, [2]Aafreen, [3]Abhishek Sharma, [4]Aryan Singh, [5]Supriya K

[1]Professor, [1,2,3,4,5]School of Computing & Information Technology, REVA University, Bangalore, India
[1]basavarajshadimani@reva.edu.in, [2]aafreen1804@gmail.com, [3]ruplusharma@gmail.com,
[4]aryansingh2599@gmail.com, [5]supriyakeeliputti@gmail.com

Article Info
Volume 83
Page Number: 4288-4293
Publication Issue:
May - June 2020

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *12 May 2020*

Abstract

Machine Learning has transformed Healthcare Domain providing more efficient, faster, smarter ways to detect and cure various diseases. Machine learning approaches are widely used for Parkinson's Disease (PD) prediction. The prediction of Parkinson's disease challenging for doctors and researchers as the symptoms of the disease are examined in middle and later middle ages where condition has progressed over time. In this paper, we will build a predictive model which can be used for early and accurate detection of the presence of Parkinson's disease in one's body. We focused on XGBoost, a new Machine Learning algorithm, based on decision trees, designed with speed and performance in mind, to improve the accuracy of PD prediction. This approach using the XGBoost algorithm obtained higher accuracy than other machine learning techniques such as Naïve Bayes algorithm, binary logistic regression, random forest and support vector machine.

Keywords**:** *Machine Learning, Parkinson's disease, XGBoost, Decision Tree Based*

1. Introduction

Healthcare domain is one the most important and largest in the world as it plays a vital role in people's life and directly affects quality of life. The number of health issues are increasing exponentially due to various factors and even though there has been a greater improvement for treatment and diagnosis in the Health sector, the need for much better methods and tools are always a necessity [14]. The improving technology and expectations always demand a better and more efficient system. If tended to accurately, through approach and practice, with a good data set and algorithm, there is a high chance to receive a scope of benefits.

Parkinson's disease is a disease of the human body's nervous system when it gradually starts breaking down and starts affecting our movement and co-ordination [1]. It may show early symptoms of tremors and overall slowing down of general movement of the body or stiffness. Our speech may also become unclear and incoherent. Parkinson's disease symptoms worsen as our condition advances over time. Spread over five stages, the symptoms may reach a stage where the face would stop showing expression and movement of the body may come to a complete halt. In other words, Parkinson's disease, if not detected early, can prove to be a huge hindrance. As it cannot be cured, regulating it through various options is the only solution to control it.[13].

The outset of machine learning has revolutionized the world. In simple words, computers have started learning on their own without being fed real time code by other developers or programmers. Through computer programs, systems have automatically started accessing data and started using it for their own benefit and learning. Data or information specifically determines how a computer system is going to learn automatically without human involvement. Data or instructions are observed, patterns are identified, past experience is gone through, and eventually spontaneous learning is done, which is changing the world around us each day.

There are various algorithms of machine learning, most of which are categorized under supervised or unsupervised[1]. In this project, Parkinson's disease was

May-June 2020
ISSN: 0193-4120 Page No. 4288-4293

tried to be automatically detected under various machine learning algorithms, viz. XGBoost, Random forest, Logistic Regression, Naïve Bayes, and SVC. The same pre-processed data was applied to the algorithms and tested. It was found that XGBoost scored an edge over all the others. We specifically tried to improve the accuracy of the same ultimately reaching an accuracy of 96.6% approximately.

The remaining part of the paper is divided as follows. Section 2 indicates literature survey followed by the motivation for this work. Section 4 is methodology. Section 5 is results and discussion. The paper is concluded with conclusion remarks and scope for further enhancement.

2. Related Work

A challenging task for radiologist is the early detection of PD. Early diagnostic helps in quick treatment plan and procedures which increases life expectancy. For this reason, we incorporate machine learning techniques for early prediction. Most of the existing works are based on diagnosing PD using artificial neural networks, resting tremor classification using multi-layer perception etc. This section describes various works carried out so far in detecting PD using ML based techniques and their limitations.

Muthumanickam S[1] et. al presents a survey of the different techniques utilized to discover, Group and Diagnose Parkinson's Disease. The author hastested Voice Samples Of Patients and has concluded that CNN, SVM and ANN are the most efficiently utilized algorithms.

Prof. Shruthi S[2] et. al proposes another system. The proposed model is trained for identification of Chronic Kidney disease, Diabetes mellitus and Heart failure. Using KNN algorithm, data Classification was done by selecting 3 different points of data from the individual disease in order to achieve an accuracy level above 50%. The experiment uses XGBoost classifier for disease predictions.

Minimum redundancy and Maximum relevance is used by Timothy J. Wroge [3], Arvind Kumar Tiwari [5] .Minimum Redundancy Maximum Relevance (mRMR) was applied to audio features that yielded an array of ranked features which were used for prepossessing data giving high accuracy.

Athanasios [10] et. al have calculated 132 dysphonia dealings. Four selected parsimonious subsets of these sdysphonia measures were selected using feature selection algorithms, and were mapped to classifiers: random forests and support vector machines. The binary discrimination problem they used reports an accuracy of 93% classification accuracy approximately. Subsequent training with SVM leads to an average performance results using the features selected by each of the four algorithms.

Zeyi [11] et. al, makes parallel implementation named GPU-GBDT for training GBDTs. They have demonstrated that GPU-GBDT can be a more efficient and cost-effective when compared to its CPU-based correspondence. Xin Gao [7] proposed an XGBoost algorithm that has shown us feature importance column sampling method for classifying images and its representations.

Akshaya and Jennifer [4] bring forward a model that diagnoses PD from data of voice recordings of Parkinson's patients and subjects without Parkinson's. They have used Genetic algorithm to select the best set of features which reduces feature vector dimension. They have experimentally tested Boosted Decision Tree model.

3. Motivation

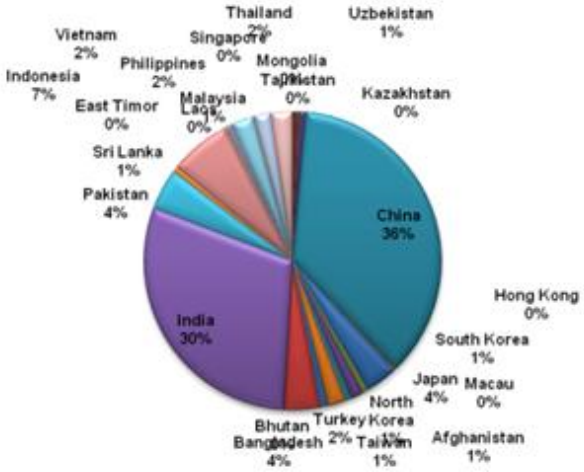


Figure 1: prevalence of Parkinson disease observed in various Asian countries

Desire to face the challenge in solving the unsolved problems, i.e., concern over practical problems initiates research. Same can be said for our project. PD has 5 stages affecting more than 1 million individuals in India every year. Figure 1. shows the prevalence of Parkinson disease observed in various Asian countries, which reveals that the disease is common in most of the countries. The maximum number of cases were observed in China (36%), followed by India (30%)[15]. It is estimated that the percentage would double in the next decade, and a major problem is the inability to effectively detect the disease at an earlier stage. Considering the horrific and unfortunate symptoms of PD as shown in figure 2, and the fact that the disease is not detected in earlier stages and by the time it is done, it's too late, early detection of Parkinson's Disease (PD) is very crucial in terms of treatment. Hence to avoid this, by using machine learning approaches, we have proposed a simple, common method in order to detect this disease on a large scale.

May-June 2020
ISSN: 0193-4120 Page No. 4288-4293

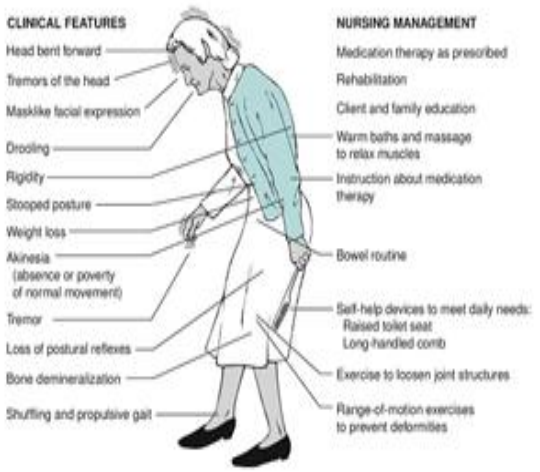


Figure 2: Symptoms and features of PD

4. Methodology

Research methodology is the detailed procedures which is used to identify, select, process, and analyze the information. The methodology section answers two main questions: How was the data collected or generated? How was it analyzed?

4.1 Dataset

UCI ML Parkinson's dataset is used for the project. The dataset has 23 columns and 197 rows. This dataset comprises of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD).The main objective of the data is to differentiate healthy people from people suffering form PD, according to "status" columns, set to 1 for PD and 0 for healthy [14].Each column is a voice measurement, and each record is one of 195 voice recording from these individuals ("name" column)as in figure 3.

Data Set Characteristics:	Multivariate	**Number of Instances:**	197
Attribute Characteristics:	Real	**Number of Attributes:**	23
Associated Tasks:	Classification	**Missing Values?**	N/A

Figure 3: UCI ML Parkinson's dataset overview

Data sets are obtained from **UC Irvine Machine Learning Repository** created by Max Little of the Oxford University along with the National Centre for Voice and Speech. The reason for choosing this data set is because the datasets are drawn from the domain, meaning that they have real-world qualities, as opposed to being synthetic. Thus, the project can be applied in real world. Setting the iteration numbers minimally can yield good results. The next step was to handle the missing values (no missing values found as it was taken from the UCI ML Repository).

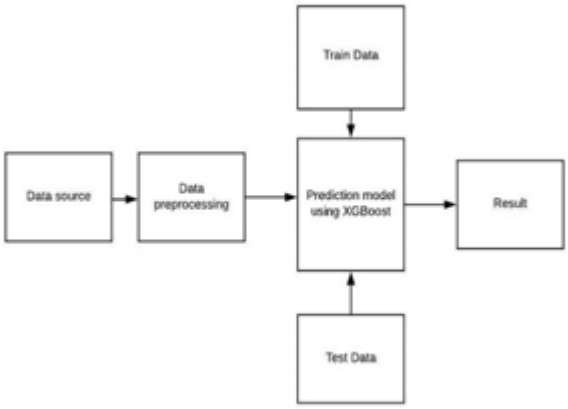


Figure 4: Flowchart of the project

4.2 Data Preprocessing

In our project, python is taken as the language of choice. We make use of various python libraries: numpy, pandas, scikit-learn and xgboost and we build a model using an XGB Classifier and compare it with other classifiers. The data is pre-processed in various ways to give new performances on the Parkinson's disease classification data. During pre-processing, we have renamed the columns of the dataset and added an id column to make it more readable and help us in later pre-processing and classification stages .We use feature selection where features like spread 1, Parkinsonian-pyramidal disease (PPD), spread 2, MDVP:Fo (Hz) Average vocal fundamental frequency, MDVP:Flo (Hz) Minimum vocal fundamental frequency are given importance.Data balancing is also performed using smote, which is a technique for increasing the number of casesin a stable way for any given dataset. Standardization is performed on the datawhich rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

4.3 Classification Using The Proposed Algorithm

Supervised learning includes an external teacher, i.e. a set of example pairs (x, y) where x and y, is needed to train the model to achieve an output for the function f: XY that matches the example in the allowed class of functions. The model creates a mapping inferred from the data. Task that fall under supervised learning is pattern recognition and regression.

The working procedure of XGBoost:

1. Firstly, boosting is a method that creates strong classifier from weak classifier by continuously adding models on top of each other one by one so that errors of previous model are corrected by the next one until the training model is totally corrected to a better one.

2. Next comes the gradient boosting part the same procedure of adding models is following but instead of putting a weight to the classifier after every step, it fits the new model to the left-out part of previous model.

May-June 2020
ISSN: 0193-4120 Page No. 4288-4293

Hence it is called gradient boosting since it updates previous models using gradient descent. Regression and classification problems both are supported by these.

3. Finally, XGBoost implements these same decision tree boosting with added speed and computational accuracy.

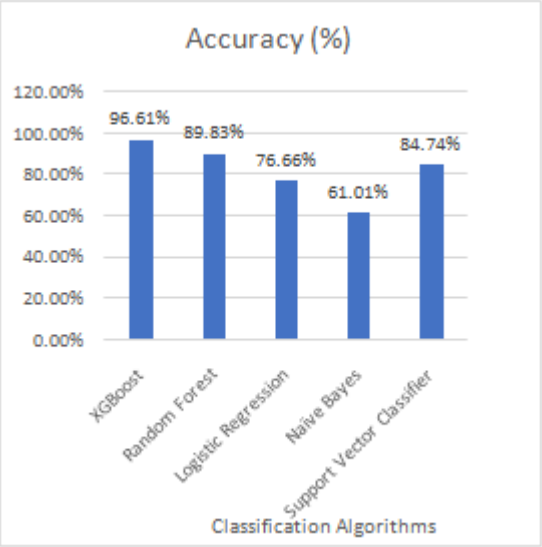
Figure 4. shows the proposed system. A different approach for classification is done using XGBoost algorithm. The training data is extracted and loaded. After training the model, test data is given to test the accuracy. The data is classified using the various ML algorithms to predict PD and show the accuracy of that prediction. We compare the various performance parameters of XGBoost algorithm with other ML algorithms and show that our proposed method is the winner.

5. Results and Discussion

The datasets will have noise. So, the pre-processing is very important. The classification data is enhanced and gives better performances now. XGBoost is an implementation of gradient boosting machines which has proven to push the limits of boosted trees algorithms as it is scalable and accurate. It includes various algorithmic advancements :Regularization**,** Sparsity Awareness **,** Weighted Quantile Sketch **,** Cross-validation . . It is a highly interpretable model. It belongs to a class of machine learning algorithms that transform feeble learners to strong ones, reducing the bias, which gives a better result than any other classification algorithms[8], proved by comparing these algorithms over few performance parameters.

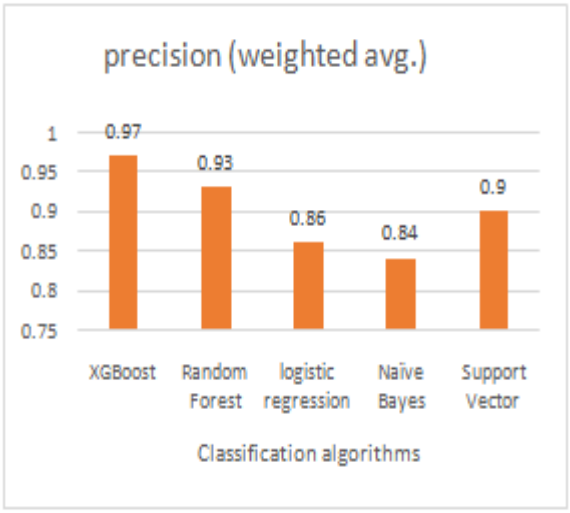
5.1 Accuracy

Accuracy (fraction of predictions the model got right)
=Number of correct predictions/Total number of predictions
(or)
Accuracy = (TP+TN)/(TP+FP+FN+TN) using the confusion matrix



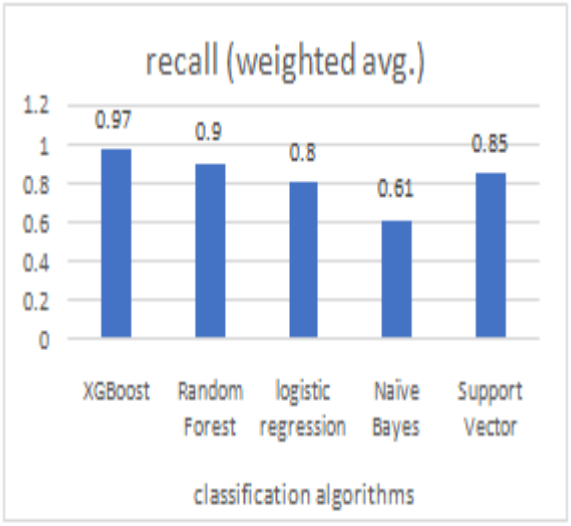
5.2 Precision

Precision (the ability to identify only the pertinent data points)
=correctly classified/(correctly classified+error classified)
(or)
Precision = TP/(TP+FP) (using confusion matrix)



5.3 Recall (Sensitivity)

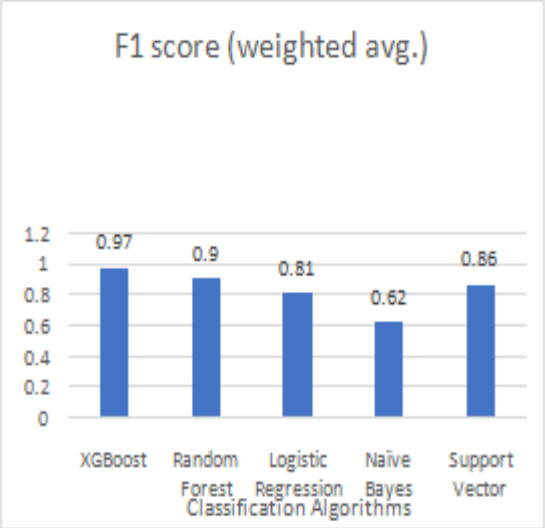
Recall (the ability of a model to find pertinent cases within a dataset)
=correctlyclassified/(correctly classified+missed classified)
(or)
Recall = TP/(TP+FN) using the confusion matrix



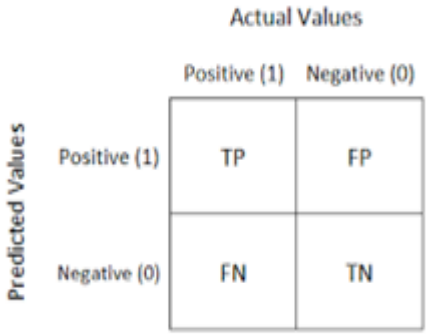
5.4 F1 Score

F1 score (F1 Score is the weighted average of precision and ecall)
= 2*(Recall * Precision) / (Recall + Precision)

May-June 2020
ISSN: 0193-4120 Page No. 4288-4293

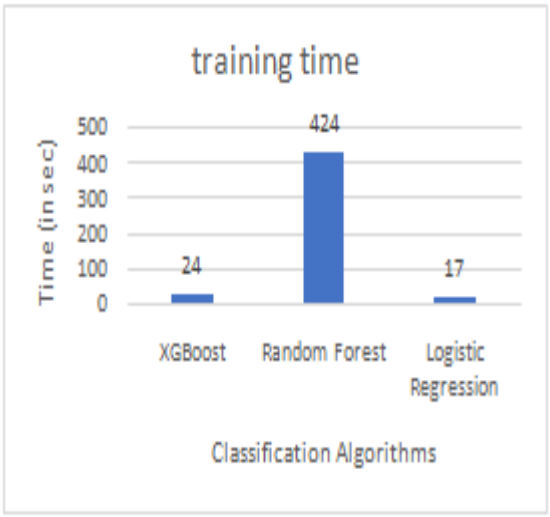


where the confusion matrix is as follows:



5.5 Training Time

Measure of time taken for processing the data(per1 million data points with 20 features)



5.6 Support

The **support** is the number of samples of the true response that lie in that class.

CLASSIFICATION ALGORITHMS	SUPPORT (WEIGHTED AVERAGE)
XGBOOST	59
RANDOM FOREST	59
LOGISTIC REGRESSION	59
NAÏVE BAYES	59
SUPPORT VECTOR	59

6. Conclusion

Diseases like Parkinson's can really prove to give a difficult time not only to the patient, but all those associated with him. Nevertheless, no one really deserves to go through all what Parkinson's has to offer with the latest developments in health around us. With technologies like machine learning, we have come quite far as a race. An early diagnosis could also help scientists in developing new and better cures for it. As put forward using this project, XGBoost with its ability to predict the disease at an early stage could change lives of many, paving the way for a better life for one and all.

Additionally, we have also compared it with various other ML algorithms over the same processed dataset and the result shows xgboost as the winner. And finally, the project has offered new performances on the Parkinson's disease classification data and provides new a perspective to study the dataset.

The work can be enhanced. A big factor in this disease is the lack awareness, and also that complication arises or be noticeable only after a certain amount of time has passed, and degeneration of neurons have already started in the subject. So, datasets with images taken earlier than the one we used, if can be obtained, can give an even better outcome. Comparing this model with other existing approaches on the premises of accuracy, efficiency and applicability on real life situations can better reveal the position of this approach.

References

[1] Muthumanickam S1, Gayathri J2, Eunice Daphne J3,"Parkinson's Disease Detection and Classification Using Machine Learning and Deep Learning Algorithms– A Survey", International Journal of Engineering Science Invention (IJESI) Conference, Volume 7 Issue 5 Ver. 1., May 2018

[2] Prof. Shruthi,Baba Bharath G,Vibha Reddy K,

May-June 2020
ISSN: 0193-4120 Page No. 4288-4293

VinuthaNagaraj, "Predicting Multiple Diseases Using Machine Learning Techniques", International Journal of Scientific Research and Review, Volume 07, Issue 05, May 2019, Page No.416, 2019

[3] Timothy J. Wroge1, Yasin ¨Ozkanca2, Cenk Demiroglu2, Dong Si3, David C. Atkins4 and Reza Hosseini Ghomi4, "Parkinson's Disease Diagnosis Using Machine Learning and Voice", IEEE in signal processing in Medicine and Biology symposium (SPMB), 2018

[4] Akshaya Dinesh, Jennifer He, "Using Machine Learning to Diagnose Parkinson's Disease from Voice Recordings", MIT Undergraduate Research Technology Conference (URTC), 2017

[5] Arvind Kumar Tiwari GGS College of Modern Technology, SAS Nagar, Punjab, India. "Machine learning based approaches for prediction of Parkinson's Disease" Machine Learning and Applications: An International Journal (MLAIJ) Vol.3, No.2, June 2016

[6] Srishti Grover, Saloni Bhartia, Akshama, Abhilasha Yadav, Seeja K. R "Predicting Severity Of Parkinson's Disease Using Deep Learning"International Conference on Computational Intelligence and Data Science (ICCIDS 2018)

[7] Xin Gao, Shaohua Fan, Xinpeng Li, Ziming Guo, Hao Zhang,Yuexing Peng, Xinping Diao, "An Improved XGBoost Based on Weighted Column Subsampling for Object Classification ". in 4th International Conference on Systems and Informatics (ICSAI), 2017

[8] Najmeh Fayyazifar, Najmeh Samadiani "Parkinson's Disease Detection Using Ensemble Techniques and Genetic Algorithm" in Artificial Intelligence and Signal Processing (AISP), 2017

[9] Y. Choi, C. Y.-I. Chiu, and D. Sontag, "Learning Low-Dimensional Representations of Medical Concepts," in AMIA Summits on Translational Science Proceedings, vol. 2016

[10] Athanasios Tsanas, Max A. Little, Patrick E. McSharry, Senior Member, IEEE, Jennifer Spielman, and Lorraine O. Ramig, "Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease", IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 59, NO. 5, MAY 2012

[11] Zeyi Wen, Bingsheng He, Kotagiri Ramamohanarao, Shengliang Lu, Jiashuai Shi, "Efficient Gradient Boosted Decision Tree Training on GPUs" in IEEE International Parallel and Distributed Processing Symposium, 2018

[12] D. Kartchner, T. Christensen, J. Humpherys, and S. Wade, "Code2vec: Embedding and clustering medical diagnosis data," IEEE International Conference on Healthcare Informatics (ICHI), Aug 2017.

[13] 'Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection', Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. BioMedical Engineering OnLine 2007, 6:23 (26 June 2007)

[14] "Parkinson's disease: clinical features and diagnosis", J Jankovic, Neurol Neurosurg Psychiatry: first published 14 March 2008.

[15] "Global Prevalence and Therapeutic Strategies for Parkinson's Disease", Faisal Nouroz, Madiha Mehboob, Sajid Khan, Tibgha Mobin,, PJCBR Vol. 2 No. 2 Jul. – Dec. 2014