International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

ELSEVIER SSRN

IJIEMR Transactions, online available on 26th Apr 2022. Link
:http://www.ijiemr.org/downloads.php?vol=Volume-11&issue= Spl Issue 04

DOI: 10.48047/IJIEMR/V11/SPL ISSUE 04/02

Title **PREDICTION OF PARKINSON'S DISEASE USING GRADIENT BOOSTING ALGORITHM**

Volume 11, SPL ISSUE 04, Pages: 13-21

Paper Authors
A.ChandraMouli, K.Lakshmi Yasaswi, Divvela Swetha, G.Naga Sowmya Sree, Kasiboyina Manisai

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

PREDICTION OF PARKINSON'S DISEASE USING GRADIENT BOOSTING ALGORITHM

[1]A.ChandraMouli, [2]K.Lakshmi Yasaswi, [3]Divvela Swetha, [4]G.Naga Sowmya Sree, [5]Kasiboyina Manisai
[1]Associate Professor, Department of CSE, PSCMRCET, Vijayawada, achandramouli@pscmr.ac.in
[2]Student, CSE, PSCMRCET, Vijayawada, yasaswikanamarlapudi@gmail.com
[3]Student, CSE, PSCMRCET, Vijayawada, dmrswetha9812@gmail.com
[4]Student, CSE, PSCMRCET, Vijayawada, sowmya6302@gmail.com
[5]Student, CSE, PSCMRCET, Vijayawada, manisai.kasiboyina@gmail.com

Abstract
Parkinson's Disease (PD) is a persistent neurological disorder, mainly influencing the 40 to 60 year age group. It affects the midbrain region called substantia nigra, which produces dopamine. Reduced dopamine concentration causes motor symptoms like tremors, bradykinesia, vocal symptoms, and non-motor symptoms like painful cramps, constipation, gastric problems, and sleeping problems in people having PD. Above 90% of the PD patients experience vocal damage that causes a symptom called, Dysphonia. As an incurable disease, it requires to predict at an early stage. With Developing technology and data, Machine Learning plays a vital role in considering the most accurate decisions at lower costs. Our ML model consists of a feature selection and classification process, which uses to predict whether a particular person is suffering from PD or not. We used a vocal dataset that contains many multivariate attributes. We use the Pearson correlation for feature selection to get the correlation based on the relation between the target attribute with others in the dataset. We trained our model with Gradient Boosting Classifier, Decision Tree Classifier, and Naïve Bayes, where Gradient Boosting gives the highest accuracy of 94.87% compared to others. We build a user interface that would take the relevant features as input from the user and displays the predicted result. The main aim of this project is to differentiate whether a person is a healthy or PD patient.

Keywords: Parkinson's Disease, Naïve Bayes, Decision Trees, Gradient Boosting, Correlation, Machine Learning.

I. INTRODUCTION

Parkinson's is a long-standing neurological disease that shows a gradual decline in the count of neurons in the midbrain called substantia nigra, a part of the basal ganglia. PD is incorrigible neurodegeneration. This illness shows symptoms of rigid muscles, tremors at rest, retarded movements, postural imbalance, cognitive impairment, and psychological problems. These signs of PD are easy to identify at their growing stage, but it is difficult to identify at an early stage. There is a neurochemical, Dopamine, responsible for the human nervous system.

Less concentration of this neurotransmitter in the primary part of the midbrain leads to PD. The dopaminergic dysfunction of the nerve cells causes different symptoms. We categorized those symptoms into the motor and non-motor signs. Parkinson's is generally considered a disease that only involves movement. The significant motor symptoms include tremor, rigidity, bradykinesia (slow movements), and postural imbalance (balancing issues). There is no compulsion that all these symptoms or signs must be present for every PD patient. In addition to motor symptoms, most people also develop other problems similar to PD symptoms. These symptoms are called non-motor symptoms that involve, Intellectual changes, Egestion, Feeling full quickly while eating, Excessive sweating, Fatigue, Hallucinations and delusion, Lightheadedness, Sleep disorders, Vision problems, and many more.

It increases the demand for advanced technologies for effective patient health management in neuro disorders. These predictive models will help to spot who are healthy and who are suffering from PD. The ML classification technique will help improve the accuracy and result of the model and the dependability of diagnosis and reduce the possible misleads, hence making PD classification more time-efficient and cost-effective. In this, we applied feature-based selection and classification to predict PD. Feature selection is an important ML approach in predicting Parkinson's disease. It helps to avoid the curse of dimensionality, helps in simplifying the model to interpret easily, reduces the training time, reduces the chances of overfitting, and enhances the generalization. For the classification of the ML model, we are using the Gradient Boosting algorithm, which is one of the dominant techniques for predictive modeling. Gradient Boosting is a top and popular technique and a neat alternative to regression and neural network classifiers. This vocal dataset consists of the collection of the multivariate attributes, and speech recordings regulated at different frequencies. Dataset consists of the records of both healthy people and PD-affected patients. For Feature-based selection, we used Pearson Correlation to get only relevant features from the dataset and apply Data Normalization using MinMax Scalar for the values not in the required range. We build a user interface to maintain communication, the user and the system give input and get the predicted results.

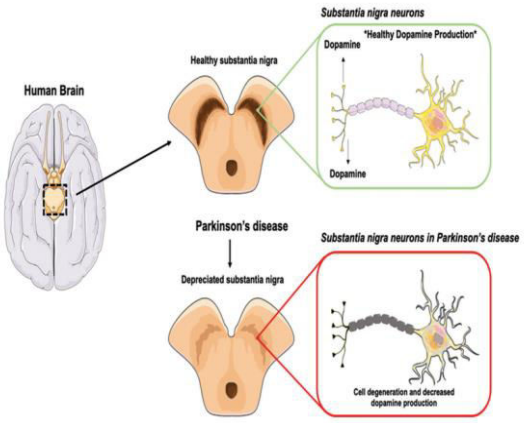

Fig 1: Substantia Nigra

II. RELATED WORK

[1] T.Sathiya et. al, proposed that

Parkinson's disease should be predicted at an earlier stage so that PD patients can get the proper treatment at the required time. PD shows up when the concentration of dopamine is less in the midbrain. They apply different ML techniques to the speech signal dataset of 30 patients at National Center located in colorado. The dataset consists of 195 biomedical voice measurements. Recursive feature elimination for choosing main features. Random forest for classification, which is a combination of various decision trees. For performance evaluation, they used accuracy, sensitivity, and specificity metrics. They got 96% accuracy for the proposed model. Their proposed system helps observe the threatening factors and gives warnings if it identifies any PD signs. In the future, if we use hybrid feature selection, it will remove irrelevant features and improves the performance.

[2] Hakan Gunduz proposed that PD is a rapidly developing illness created by the deprivation of dopamine in the middle brain region called substantia nigra and affects vocal cords at an early stage resulting in pronunciation modification. The author's proposed method took the lead on the major functionalities of the variational autoencoder. For feature selection, they used a filter-based approach. They used the speech signal dataset available on Kaggle. For the classification, the proposed SVM Algorithm. To enhance the method and get cavernous features, they apply multi-kernel SVM classifiers. The Feature Selection metrics are relief and fisher score. For model evaluation, they used accuracy, f-measure, and Matthews correlation. Finally, the proposed model got 95% accuracy. For future work, the author suggested that we have to use deep learning techniques and sensors to extract features intensively and can get better accuracy.

[3] Ramadugu Akhil et.al prompted, that there are many symptoms of Parkinson's disease. Vocal issues, tremors, mobility difficulties, and psychological issues. They applied multiple classification algorithms like Logistic Regression, KNN, SVM, and Gradient Boosting to train the dataset. Feature selection is used to get better accuracy. They used Principal Component Analysis (PCA) for the feature selection process and reduced it to 24-feature datasets from the 90-feature dataset. They used different Evaluation Techniques like Accuracy Score, FI Score, Confusion Matrix, and Receiver operating characteristic curve. They got the highest accuracy of 93% for SVM, and 89% for Random Forest. At last, they conclude that in the future we can use better models using deep learning.

[4] Zehra Karapinar Senturk proposed that Parkinson's disease is caused by the disarranging of the neurons in the brain. They used a multivariate vocal dataset from UCI Machine Learning Repository. In their system, they used two methodologies. One is feature selection, and the other is the classification process. The author used RFE for feature selection which eliminates unwanted attributes. The author applied ANN, SVM, and the combination of classification and regression called CART

algorithms on training data. Finally, SVM with RFE gives better performance with an accuracy of 93.84% when compared to others. According to the author, we can get better outputs with feature selection, and this method is more useful when dealing with large datasets.

[5] Yulianti et.al, proposed that Parkinson's is a neurological disease that affects the brain and causes different problems in the human body. So it is very crucial to prevent and cure these types of illnesses at an early stage. They used a voice dataset from UCI Machine Learning Repository. The dataset contains unimportant and redundant features which reduces model efficiency. So they applied feature selection for better performance. They used a decision tree algorithm for the classification. They calculated performance metrics like accuracy, confusion-matrix, and roc curves. Then the model gave 64.17% accuracy for the decision tree,71.74% accuracy for Forward Selection in the decision tree, and 69.42% accuracy for backward selection in the decision tree. They suggested that in the future bagging techniques will reduce misclassification and improve accuracy.

III. PROPOSED SYSTEM

This paper focuses on a Machine Learning system that uses parameters like Problem-Solving skills, Technical Activities, and Non-Technical Activities to predict Parkinson's Disease. In our system, we use a Multivariate Vocal dataset is used as the sample dataset, which is taken from Kaggle, and pre-processing is done it. We are using MinMax Scalar for Data Normalization, and Pearson Correlation for Feature Selection to remove irrelevant features. The Classification uses Decision Trees, Naïve Bayes, and Gradient Boosting Algorithms for training the data making predictions. Gradient Boosting Classifier gives the highest performance when compared to other Algorithms. Our System contains User Interface that takes vocal features as input from the user and displays the predicted results using the model.

To develop Parkinson's Prediction System, we will perform the below steps.

- Data pre-processing
- Classification
- Final prediction

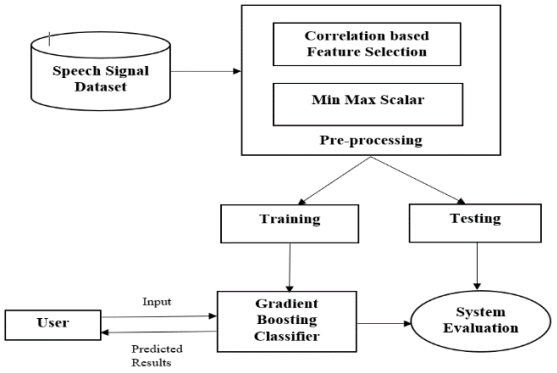

Fig 2: System Architecture

A. Data Preprocessing

Data pre-processing is the method of changing original data into an understandable format for Machine Learning. Pre-processing has been shown to improve the effectiveness of Machine Learning Systems in previous research. In this system, the proposed model uses Pearson Correlation and MinMax Scalar techniques. Because we can't work with raw

data, this is a crucial phase in the data mining process. Data cleaning has an impact on the efficiency of Machine Learning Systems. During the data cleaning process, the suggested systems discovered two irrelevant features namely, 'MDVP: Fhi(Hz)', and 'NHR' which have a negative influence on the system accuracy.

B. Preprocessing using Pearson Correlation

Feature Selection is an important step in data pre-processing. It improves the predictive power of model by choosing only relevant features and removing the redundant, irrelevant features. Correlation is used to know the dependency between the variables. The proposed System uses Person Correlation which tells how strongly every feature is linearly dependent on every other feature in the dataset. It returns the value in the range of -1 to 1 where '-1' specifies a powerful negative relationship, '0' specifies no relationship and '1' specifies powerful positive relationship between attributes or features.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

C. Preprocessing using MinMax Scalar

Data transformation is also a key pre-processing step, which is applied as MinMax Feature Scaling in the proposed system. MinMax Scalar shrinks the data within the given limit. It transforms the values to the range of (-1,1) without changing the shape of the original distribution. Usually, it takes the range from 0 to 1 but we are applying the range from -1 to 1 since we are having negative values in the dataset.

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

D. Classification

Classification is about designating the given dataset and predicting the target variable. It is supervised learning which specifies the given input into a particular category based on class labels. Classification is one of the predictive modeling methods which is used to predict the class label for the given sample of data. After training the model with specific classifiers if we give new input then it will be the output that consists of a specific class. We have different types of classification algorithms in machine learning. We can use one algorithm at a time and we can combine different classifiers called meta classifiers to construct a model.

Among various classification algorithms, here the system uses classification techniques as:

- Decision Trees Classifier
- Naïve Bayesian Classifier
- Gradient Boosting Classifier

i) Decision Trees

Decision Trees based prediction is one of the robust and familiar techniques which is commonly used for classification and regression in machine learning. A decision tree is nothing but a hierarchical or normal tree structure consisting of branches and nodes (root node, internal nodes, leaf

nodes), where leaf nodes represent class labels for the given classification problem. We can apply this technique to both linear and non-linear datasets. It is mainly used to make decisions by selecting features to classify the given problem and it is also used as feature engineering by removing irrelevant features. It is a base model for many ensembling algorithms like boosting, Bagging, and Random Forest. We use different metrics like entropy and information gain for building this decision tree.

ii) Naïve Bayes

Naïve Bayes in Machine Learning is considered a probabilistic model in supervised learning used in different use cases. It is one of the fast and furious techniques where we can solve machine learning problems effortlessly and smoothly in less time. This algorithm applies to both binary and multi classified datasets. Naïve Bayes is called Idiot Bayes because each hypothesis calculation is clear to make it tractable. There is an assumption, that each feature is independent and makes an equal contribution to the outcome. It is to predict if the weather will be good or bad. Doctors can diagnose their patients by using the information given by the classifier.

iii) Gradient Boosting

Gradient Boosting is one of the popular Boosting algorithms, this is Machine learning Boosting type. The main benefit of this algorithm is a complex reduction and time-saving. We apply this method to both classification and regression. It also improves complex problem-solving approaches in machine learning. It strongly depends on the prediction, that the next model will reduce the errors when combined with the previous one. This model is a strategy that combines multiple simple weak learners into a composite model. With other added simple models, this model will become a strong predictor. Here Gradient Boosting consists of two main functions. They are optimization function and loss function which are helpful to optimize the problem and improve accuracy by changing weights to data points in every iteration. It combines all weak learners into strong learners to increase effectiveness. Usage of different types of loss functions results in flexibility, applied for regression, multi-class classification, and many more. Gradient Boosting is a stage-wise additive model that adds weak learners to the learning process. The contribution of the learners depends on the Gradient descent optimization process, and calculation depends on the overall error rates reduced in the strong learners. It trains the remains of the models, which acts as an alternative to giving more importance to the misclassified observations.

IV. PERFORMANCE METRICS

In the proposed system, the Naïve Bayes algorithm was obtained with 74.35% accuracy, and the Decision Tree algorithm was obtained with 84.6% accuracy. To get better accuracy we use the Gradient Boosting algorithm, which gives 94.87% accuracy which is higher when compared to other algorithms.

The performance evaluation metrics are

calculated are

A) Accuracy

Accuracy is the value or percentage that tells how correctly or accurately the output is coming. In machine learning, accuracy means the value of measurement to determine the correctness of the model for the given training dataset. The formula of accuracy is given below:

Accuracy = total no.of predictions / total no.of samples

B) Precision

Precision is one of the important performance measuring a parameter in machine learning. Precision means the quality or fact of being accurate i.e, positive predictions. We can get the precision value by dividing all true positive values by all positive values.

Precision = tp / (tp+fp)

Where 'tp' is true positives and 'fp' is false positives.

C) Recall

The recall is one of the performance metrics which is used in classification models. It is the value that tells about the correct hits. We can get the value of recall by dividing the true positives by the summation of true positives and false negatives.

Recall = tp / (tp+fn)

Where 'tp' is true positives and 'fn' is false negatives.

D) Confusion Matrix

Confusion matrix is frequently used system evolution technique. It is mainly used for evaluating the performance of all types of classification problems. It is also known as error matrix because it gives the errors in the model for given dataset. The matrix is represented as 2x2 if it is a binary classification and it is represented as nxn matrix where n is the number of target classes. It has mainly two dimensions actual values and predicted values.

The output of the confusion matrix is:

```
[[ 5  2]
 [ 0 32]]
```

Table 1: Comparative table of different algorithm results

S.NO	Algorithm	Dataset	Accuracy	Recall	Precision
1	Naïve Bayes	Parkinson	74.35%	0.75	0.92
2	Decision Trees	Parkinson	84.61%	0.88	0.93
3	Proposed System	Parkinson	94.87%	1.00	0.94

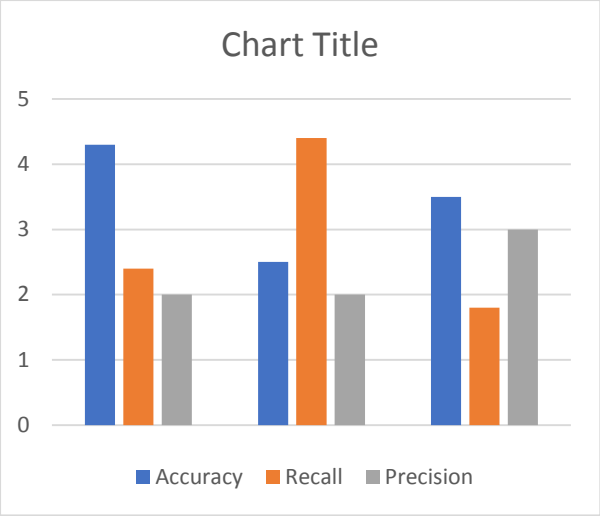

Fig 3: Comparison Graph for Performance

V. CONCLUSION

In our proposed system, we build a Gradient Boosting model to classify Parkinson's Disease (PD). The vocal dataset taken from Kaggle is used to form training and testing data with the 80:20 test ratio. data testing was most accurate in Gradient Boosting with 94.87% of accuracy among all classifiers. We used Pearson Correlation for feature selection which reduces Overfitting, Training Time and improves Accuracy by selecting only relevant features. And for Classification we used Gradient Boosting Algorithm, Our model gives better accuracy when compared with Naïve Bayesian and Decision Tree. We also applied Normalization for a few attributes in the dataset that don't have values in the range of -1 to 1. People with PD mostly have less motor function and capacity, as well as muscle and bone loss when their brains don't have enough neurotransmitters. These people, too, have to deal with the difficulties of normal aging. This lowers the quality of life, makes people afraid of falling, and makes them choose to stay at home. Parkinson's prediction model helps in predicting whether they have Parkinson's disease or not, this will help in controlling it from further advances.

REFERENCES

[1] T. Sathiya, R. Reenadevi, B. Sathiyabhama. (2021). Random Forest Classifier-based detection of Parkinson's disease. Annals of the Romanian Society for Cell Biology, 2980 –. Retrieved from https://www.annalsofrscb.ro/index.php/journal/article/view/4912

[2] Hakan Gunduz, An efficient dimensionality reduction method using filter-based feature selection and variational autoencoders on Parkinson's disease classification, Biomedical Signal Processing, and Control, Volume 66,2021,102452, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2021.102452 (https://www.sciencedirect.com/science/article/pii/S1746809421000495)

[3] Matthew P. Adams, Arman Rahmim, Jing Tang, Improved motor outcome prediction in Parkinson's disease applying deep learning to DaTscan SPECT images, Computers in Biology and Medicine, Volume 132,2021,104312, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2021.104312 (https://www.sciencedirect.com/science/article/pii/S0010482521001062)

[4] Zehra Karapinar Senturk, Early diagnosis of Parkinson's disease using machine learning algorithms, Medical Hypotheses, Volume 138,2020,109603,

ISSN03069877, https://doi.org/10.1016/j.mehy.2020.109603 (https://www.sciencedirect.com/science/article/pii/S0306987719314148)

[5] Yulianti & Syapariyah, A & Saifudin, Aries & Desyani, Teti. (2020). Feature Selection Techniques to Choose the Best Features for Parkinson's Disease Predictions Based on Decision Tree. Journal of Physics: Conference Series. 1477. 032008. DOI:10.1088/1742-6596/1477/3/032008

[6] V. Sharma, S. Kaur, J. Kumar, and A. K. Singh, "A Fast Parkinson's Disease Prediction Technique using PCA and Artificial Neural Network," *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, 2019, pp. 1491-1496, DOI: 10.1109/ICCS45141.2019.9065876.

[7] Nagasubramanian, G., Sankayya, M. Multi-Variate vocal data analysis for Detection of Parkinson disease using Deep Learning. *Neural Comput & Applic* **33,** 4849–4864 (2021). https://doi.org/10.1007/s00521-020-05233-7

[8] Matthew P. Adams, Arman Rahmim, Jing Tang, Improved motor outcome prediction in Parkinson's disease applying deep learning to DaTscan SPECT images, Computers in Biology and Medicine, Volume 132,2021,104312, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2021.104312 (https://www.sciencedirect.com/science/article/pii/S0010482521001062