PARKINSON'S DISEASE DETECTION USING MACHINE LEARNING

Shreevallabhadatta G[1], Suhas M S[1], Vignesh[1], Manoj C[1], Rudramurthy V C[2], Bhagyashri R Hanji[3]

[1]Students, Department of Computer Science and Engineering, Global Academy of Technology
[2]Asst.Professor, Department of Computer Science and Engineering, Global Academy of Technology
[3]Head of the Department, Department of Computer Science and Engineering, Global Academy of Technology

---***---

Abstract - *In this Global era, Technology plays an important part in our lives, considering our Lifestyle, Healthcare and maintaining resources and assets. In the field of HealthCare, technology has been growing each day in order to counter different diseases and their symptoms emerging in the present world. One such disease is Parkinson's Disease. Parkinson's Disease is a brain neurological disorder. It causes tremors in the body and hands, and also stiffness in the body. At this moment, there is no proper cure or treatment available. Only when the condition is detected early, or at its onset, is treatment possible. These will not only lower the cost of the sickness, but they may also save lives. As a result, a project called "Parkinson's Disease Detection Using Machine Learning Technologies" was launched in try to diagnose the disease at an early stage. Parkinson's disease is a neurological disease that affects the brain's dopamine-producing neurons and progresses over time. As a result, various machine learning techniques and Python libraries are employed in order to develop a model capable of reliably detecting the presence of disease in one's body. The current models rely on image or audio analysis to diagnose disease, encouraging the development of a new model that uses both.*

Key Words: **Parkinson's disease, deep learning, ensemble learning, early detection, premotor features, features importance.**

Five major sections of the paper are:

1. Abstract
2. Introduction
3. Research Elaboration
4. Results or Finding
5. Conclusion

2. INTRODUCTION

This entire Data Processing process can be automated and effective by using Machine Learning algorithms, mathematical modelling, and statistical expertise. Graphs, movies, charts, tables, photos, and a variety of other formats can be generated as a result of this entire process, depending on the task at hand and the machine's requirements. Early detection is currently the best way to tackle Parkinson's disease. To protect neuron integrity and reduce the progression of Parkinson's disease, it is critical to diagnose it early. Various Machine-Learning Techniques and Algorithms can assist patients in getting early medication or treatment and in prominent journals to finish their grades. Furthermore, published research work carries a lot of weight when it comes to getting accepted into a prestigious university and improving medical standards. Let's have a look at a few machine learning applications in the healthcare industry, as seen in Figure 1.

Figure1: Application in Healthcare Sector

Machine Learning: Machine learning (ML) is the study of computer algorithms that can learn and develop on their own with experience and data.

Parkinson's Disease (PD): Parkinson's disease (PD) is the most prevalent movement disorder caused by neurodegeneration. Degradation of dopaminergic neurons is a feature of this condition.

3. RESEARCH AND ELABORATION

3.1 RELATED WORK

Shrihari K Kulkarni, K R Sumana, the researchers in [1] used Decision Tree, Logistic Regression, and Naive Bayes, Deep Learning algorithm like Recurrent Neural Networks (RNN) by predicting the Performance Parameters to build the model. Machine learning approaches will be used to construct prediction models that can differentiate early PD from healthy normal using the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS). For Subject and Record Validation, Logistic Regression, Random Forests, and Support Vector Machine were employed.

Drawback of this paper were, Data Collection techniques are weakly regulated, resulting in unreliable results such as out-of-range or non-existent This Model purely relies on Evaluation of Motions which is not the only source of Data available on the Disease-bearers or Healthy Citizens.

Yatharth Nakul, Ankit Gupta, Hritik Sachdeva, the researchers in [2] used Supervised Learning Algorithms such as Random Forest, Support Vector and Naïve-Bayes are also compared. Confusion matrix was used for accuracy checking and different Classification methods were used.ML classification technique will improve the accuracy and reduce possible loopholes. Hyper parameter tuning is used to achieve the maximum accuracy. Achieved maximum accuracy of 98.30% using the K nearest neighbor classification

The main drawbacks are Delay in Results derived and Output Progression is slow and Best Proposed Methodology used gives Higher error rate when Confusion Matrix is plotted.

SGD (Stochastic Gradient Descent) is utilized for training data models, according to Wu Wang, Junho Lee, Fouzi Harrou, and Ying Sun of [3]. The FNN (Feed-Forward Neural Network) is put into action. The sensitivity of the linear discriminate analysis approach utilized is the best, which means it has the best likelihood of distinguishing a real patient. The proposed deep learning model had a 96.45% accuracy rate. This is owing to the deep learning model's favorable capabilities in learning linear and nonlinear features from PD data without the requirement for hand-crafted feature extraction.

The biggest drawback is that Deep Learning is frequently employed as a Blackbox algorithm, the trained neural networks are difficult to evaluate. Theoretically, it's difficult to comprehend how deep learning generates good results.

Support Vector Machine (SVM), Feedforward Back-Propagation Based Artificial Neural Network (FBANN) And Random Tree (RT) Classifiers, Binary Logistic Regression, Linear Discriminant Analysis (LDA), Convolutional Neural Network (CNN)Deep Belief Network (DBN) Technique Deep Neural Network Classifiers were used by Muthumanickam S, Gayathri J, Eunice Daphne J, the researchers in [4]. It has a higher level of accuracy than a deep neural network. Linear regression is simple to comprehend. It can be tweaked to prevent overfitting. The sgd command can be used to update linear models. The Algorithm and the Outputs of Binary Logistic Regression Have a Good Interpretation.

The main drawback is that there are a lot of data to train with, and the computing expenses are higher. In non-linear relationships, linear regression fails miserably. They aren't adaptable, difficult, or quick.

The researchers in [5], Timothy J. Wroge, Yasin Ozkanca, Cenk Demiroglu, and Dong Si, employed the VAD algorithm (Voice Activation Detection Algorithm) to clean the dataset. Cross Validation is performed using a decision tree and a Support Vector Machine. Neural Networks developed from Keras and Tensor Flow were also utilized. Machine learning architectures based on non-invasive vocal biomarkers can be used to diagnose and forecast disease. For noisy and high-dimensional data, machine learning classifiers are effective.

Accuracy of this Particular model is very less comparatively. Algorithm's Performance is limited considering it is only Clinician's Data.

Jayashree R. J, Ganesh S, Karanth S.C, Lalitha S has deeply explained how spectral features ex. Spectral Contrast, STFT and temporal features ex. Zero Crossing rate are extracted and classification done using XGBoost and Classifiers including Random Forest and Regression like Logistic Regression. It has Several Advantages including the Real Time Speech Analysis which perfectly shows how noise and other factors affect Parkinson's Disease Detection. It depicts better Accuracy and a different Characteristic approach on proving how noise affects in the Prediction of Parkinson's Disease. Its biggest drawback is it is based on a limited dataset, if more data was available, a more practical approach could be designed. Limited Classifiers are used and Analysis of Features are constrained to just eight features which is very limited comparatively.

3.2 SYSTEM DESIGN

The Proposed Architectures has totally 4 parts where it involves Dataset i.e., Data Acquisition, Feature Extraction, Classification and Output production as shown below in Figure 2.

- Data Acquisition involves acquiring all the data available which involves voice samples of People which contains noise or noiseless Features.
- Feature Extraction involves Voice analysis based on its Features i.e. MDVP.Jitter etc.
- Classification involves processing the given features of dataset using different classifiers such as SVM, XGBoost and Classification, Regression using Random Tree Classifier and Logistic Regression.
- Final Output is Produced for Prediction of the Disease involving individual Interests.

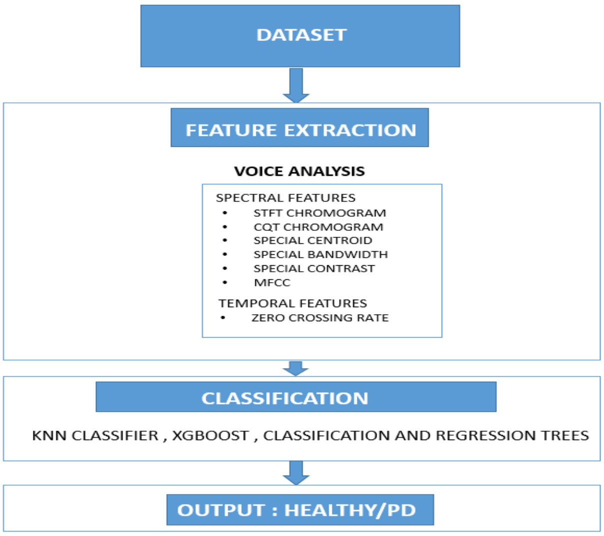

Figure 2: System Design

3.3 DATAFLOW DIAGRAM

The Data flow diagram explains the basic flow data through various steps of Parkinson's disease Detection.

• In the first step we will collect data like from the patient with different medical equipment.

• The collected data is now sent process for training the data which will be classified using different ML algorithms.

• After classification the data into Train and Test datasets, the datasets will be sent to predict whether the patient has the disease or not.

The below figure3 shows the dataflow diagram for the proposed solution for the detection of Parkinson's Disease.

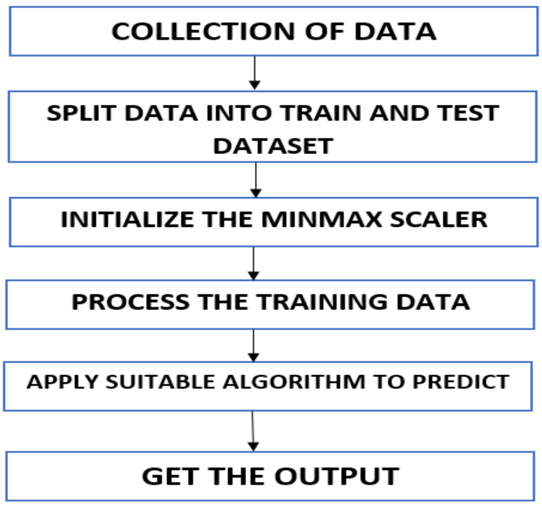

Figure 3: Dataflow Diagram1.0

• The Figure4 gives the combined analysis of both voice and image analysis to detect Parkinson's disease.

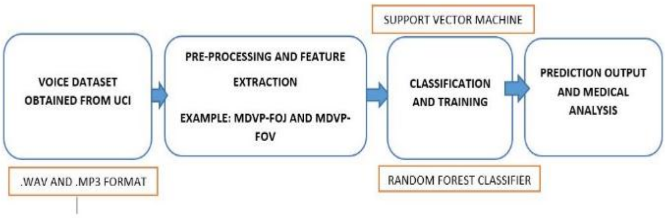

Figure 4: Data flow diagram2.0

3.4 CLASS DIAGRAM

The Fig 5 explains the class diagram of Parkinson's, detection.

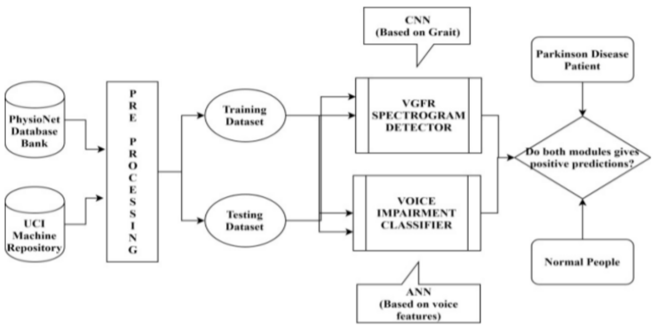

Figure 5: Class diagram

3.5 MODULES

• MODULE 1 – Dataset Extraction.

Functionality: Importing different modules for data analysis, data cleaning, model building. Importing dataset from fixed folder or directory. Voice dataset is the input. Importing dataset for data analysis and cleaning. Assigning a data frame variable, the dataset for analysis. Fetch the features and targets from the data frame using pandas. Dataset involves acquiring all the data available which involves voice samples of People which contains noise or noiseless and SPECT images available. Figure 6 explains importing different modules for data analysis, data cleaning, model building.

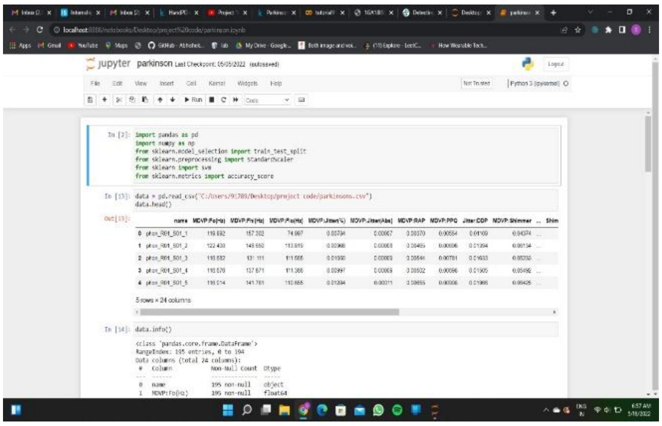

Figure 6: Importing dataset

• MODULE 2 – Dataset Analysis and Dataset Cleaning

Here the imported voice dataset is chosen for analysis. The unwanted data will be removed and some of the null data will be added with a mean value. Imported voice dataset having features like jitter, shimmer, etc. is the input. Data in columns of the dataset contains null value which should be filled certain values so that we can build error free model. Some of the columns in the dataset are not required for the classification, so we can remove unwanted columns. Data in columns of the dataset contains null value which should be filled certain values so that we can build error free model. Figure 7 explains data analysis.

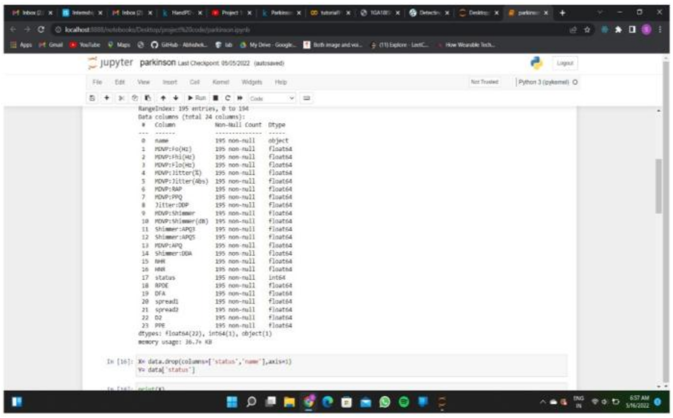

Fig 7: Data analysis

Figure8.1 and 8.2 explains data cleaning.

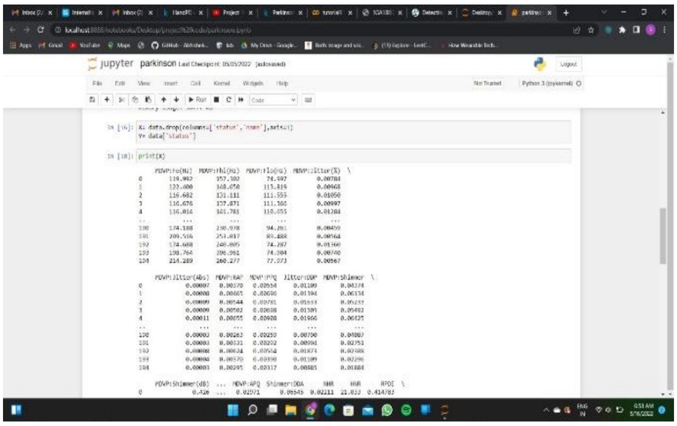

Figure 8.1: Data cleaning

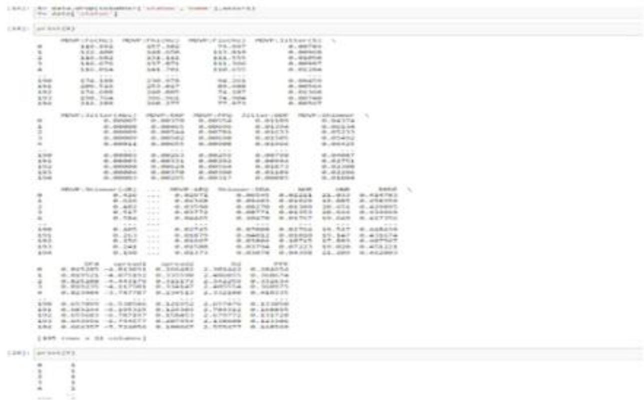

Figure 8.2: Data cleaning 2.0

• MODULE 3 – Data Splitting into Training and Testing Datasets

Splitting the cleaned dataset into Training and Testing datasets for model building. Split the dataset into training and testing sets where twenty percent data for testing purpose. Input is voice datasets. Output of this module is the two datasets, randomly distributed to training and testing datasets each containing both input features and target value. We used test_train_split from sklearn. model.selection to split the dataset into eighty percent training dataset and twenty percent test dataset. Figure 9 explains Dataset Splitting and Decision Trees plotted to analyze the best valued approach for specified output.

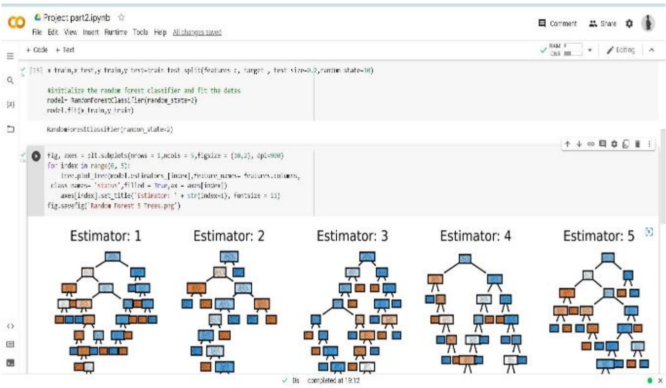

Fig 9: Dataset splitting

• MODULE 4 – Model Building

Testing the model with unseen dataset is the primary step. Input for this step is Trained model and Unseen Dataset. Output of this step is to check the model accuracy with various classification metrics. Add the testing data to the trained model. Use classification metrics like Accuracy Score, F1 Score, Precision Score and Recall Score to check the classification accuracy of the trained model. In the next step we are going to add our training dataset to the built model. We are going to check the accuracy of both testing and training dataset. Here we have used SVM (support vector machine) or Random Forest Classifier to classify whether the patient has Parkinson's or not. Figure 10 explains model building where Figure 11 shows the Final Output depicting all the Classification Metrics and Detection of the Disease for a particular patient.

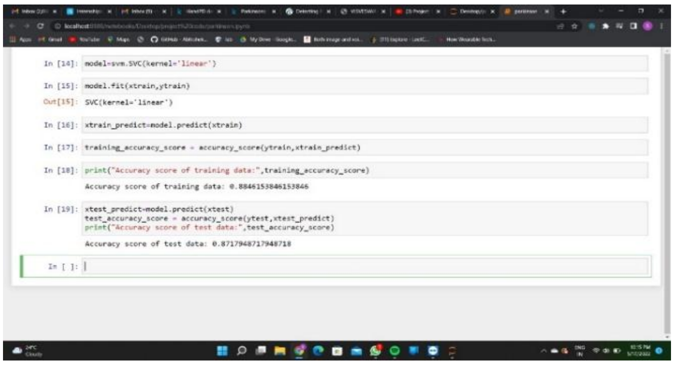

Fig 10: Model building

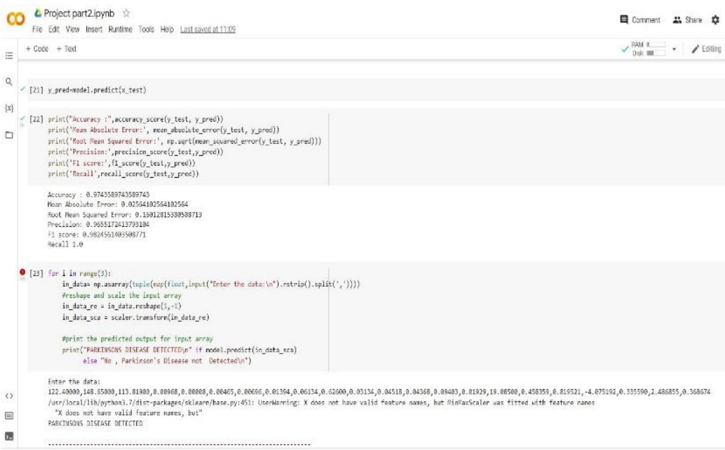

Fig 11: Final Output and Prediction

3. RESULTS AND FINDINGS

Performance Metrics like Accuracy Score, Mean Absolute Error, Root Mean Absolute Error, Precision, F1 Score, and Recall Score, among others play a prominent part in the outcomes of any Machine Learning project. Gathering all of these factors and metrics is critical since it allows us to assess the model's strengths and weaknesses. When producing predictions in novel scenarios and other Sequenced Oriented Cases, model performance is critical for machine learning. We plotted the Performance Metrics against certain Models after comparing all of the Models based on our Dataset.

1. Accuracy Score: It is the most frequently applied parameter in all models; it is the ratio of True Positives and True Negatives to all Positive and Negative Observations. Figure 15 shows a graphical depiction of Accuracy Scores for the various models we examined in this project.

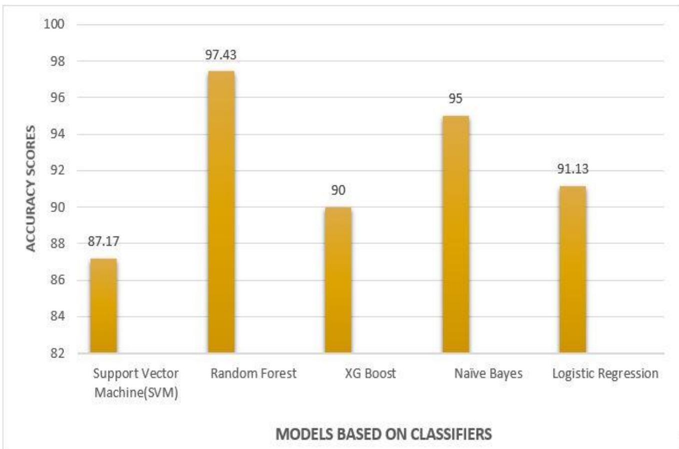

Fig 15: Accuracy score

2. Precision Score: Measures the percentage of labels and predictive values that are positively predicted. When the classes are imbalanced, Precision Score is a useful measure of prediction success. Mathematically, it is the ratio of true positive to the sum of true positive and false positive. The graphical depiction of Precision Score against the various models we utilized in this project is shown in Figure12.

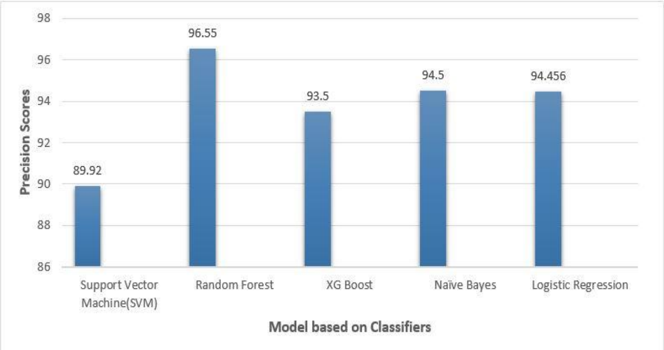

Figure 12: Precision Score

3. F1 score – Overall Incorporator Inculcating both Precision and Recall. Often used for Optimization based on Models. Figure13 gives the F1 Score graph of various models of classification.

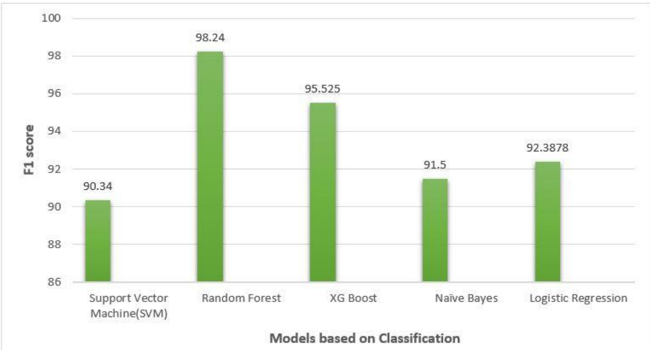

Figure 13: F1 Score

4. Recall Score: It is the accurate enumeration of real values from actual positive values. This Score indicates how significant this component is in optimizing Required Output. The Figure14 offers a graphical analysis of how different models calculate recall scores.

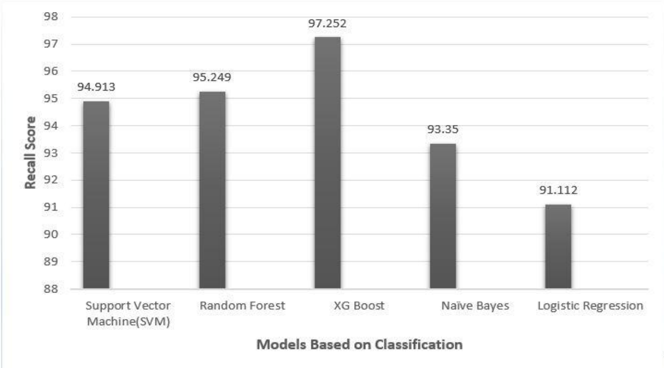

Figure14: Recall Score

It can be Inferred from the Above Tables such that Random Forest outperforms all the Other Models implemented with Accuracy Score around 97.43 %, Precision Score around 96.55%, F1 Score around 98.24%, whereas XG Boost lead the Recall Score with 97.252%. Hence Considering all the Performance Metrics, Random Forest is the most obvious choice considering the fact that it has outperformed and can be trusted on such a problem which doesn't allow even a small gap for error considering it is a part of a field like Health Care.

4. CONCLUSION

Parkinson's disease is a brain disorder that affects the central nervous system (CNS), and there is currently no cure for it unless it is diagnosed early. Late detection results in no therapy and death. As a result, early detection is critical. We used machine learning algorithms such as SVM (Support Vector Machine), Decision Tree, Random Tree Classifier, and Neural networks for early disease detection because they are known for their efficiency and quick retrieval. Most importantly, speech processing has a lot of potential in terms of Parkinson's disease detection, classification, and diagnosis. We expect that more Machine Learning-based technologies and medical techniques will be available soon to save people from this disease.

REFERENCES

[1] Shrihari K Kulkarni1, K R Sumana2,"Detection of Parkinson's Disease Using Machine Learning and Deep Learning Algorithms" International Journal of Engineering Science Invention (IJESI) ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726 VOLUME 8 ISSUE:8, (AUG 2021)-Page No :1189-1192.

[2] Yatharth Nakul1 , Ankit Gupta2 , Hritik Sachdeva3,,," Parkinson Disease Detection Using Machine Learning Algorithms" International Journal of Science and Research (IJSR) ISSN: 2319-7064 SJIF (2020): 7.803 Volume 10 Issue 6, June 2021-Page no 314-318.

[3] Wu Wang1 , Junho Lee2 , Fouzi harrou3 and Ying sun4,,," Early Detection of Parkinson's Disease Using Deep Learning and Machine Learning" IEEE ACCESS Digital Object Identifier 10.1109/ACCESS.2020.3016062 Volume 8,2020-Page no 147635- 147646.

[4] 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) ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726 www.ijesi.org ,Volume 7 Issue 5 Ver. 1, May 2018 || PP 56-63.

[5] Timothy J. Wroge1 , Yasin Ozkanca 2 , Cenk Demiroglu3, Dong Si4 , David C. Atkins5 and Reza Hosseini Ghomi6,," Parkinson's Disease Diagnosis Using Machine Learning and Voice", Conference: 2018 IEEE Signal Processing in Medicine and Biology Symposium||DOI:10.1109/SPMB.2018.8615607.

[6] Jayashree R.J., Ganesh S., Karanth S.C., Lalitha S. (2021) Automatic Detection of Parkinson Speech Under Noisy Environment. In: Thampi S.M., Gelenbe E., Atiquzzaman M., Chaudhary V., Li KC. (eds) Advances in Computing and Network Communications. Lecture Notes in Electrical Engineering, vol 736. Springer, Singapore. https://doi.org/10.1007/978- 981-33-6987-0_16.