Detection of Parkinson's disease with ensemble learning and neural network using vocal biomarker

Kanika Gandhi
Electrical and Computer Engineering
Western University
London, Canada
kgandh27@uwo.ca

Karanpartap Singh Aulakh
Electrical and Computer Engineering
Western University
London, Canada
kaulakh4@uwo.ca

Kiranjot Kaur
Electrical and Computer Engineering
Western University
London, Canada
kkaur89@uwo.ca

***Abstract*—Parkinson's disease is a multifaceted progressive neurological disorder, affecting about 7 to 10 million people worldwide, which makes it an important disease on which extensive research should be laid. A mechanism should be designed so that it can be detected early and can be stopped from progressing any further in the patients. Though this disease can't be fully cured, medications, surgeries, and physical therapy specific to this disorder's precautionary measures can be figured out so that its development is stopped from worsening. This paper offers a novel combination of ensemble learning with neural network, that has been obtained post a competitive analysis between- Bagging, Boosting and Stacking, to predict the presence of the Parkinson's disease from extracted vocal features. Although, the XGBoost classifier competes closely with the stacking model, the latter, when combined with a single layer perceptron as it's meta-model, performs impeccably on unseen data, yielding an accuracy of 98.3%.**

Keywords—Ensemble, Bagging, Boosting, Stacking, Parkinson's, Random Forest, XGBoost, Naïve Bayes, Support Vector Classifier, Neural Network.

I. Introduction

Parkinson is a progressive brain disorder that affects the nervous system and the parts that are controlled by nerve cells. The major cause of Parkinson's is the impairment of nerve cells in the basai ganglia, an area of the brain that is associated with movements. Nerve cells are actively involved in producing dopamine but when they die/ get impaired, they produce less dopamine causing issues in movement. Till date it's unclear what causes nerve cells to get impaired or die. Parkinson's is associated with an array of life-impacting symptoms such as impaired balance and coordination, muscle stiffness, tremors in the body, depression, swallowing problems, urinary problems, and skin problems to name a few. If the disease is given time to progress, the patient might also encounter problems walking and talking. Not only this, the patient experiences mental and behavioral changes, sleep problems, memory difficulties, fatigue, and depression.

Early diagnosis and early detection are the only two ways to help improve the lives of patients ridden with this life-threatening disease. With action at an early stage, many life changes can be included which in turn would beneficially reduce the torturing symptoms of this disease.

Traditionally, the diagnosis of this disease solely depends on the examination and evaluation by a clinician. Clinical assessment alongside an in-depth analysis of symptoms are pre-requisites for a doctor to confirm whether the patient suffers from Parkinson's or not. However, there are more than just a few limitations of using the traditional approach.

Firstly, this approach is highly susceptible to human error due to the manual diagnosis done by the physicians. Based on personal judgement, the interpretation of the various symptoms differs from doctor to doctor and can lead to inaccuracy in diagnosis or missed diagnosis, which in turn can pose as a high risk. Secondly, the imaging techniques used to perform clinical assessment (such as brain scans or MRI) and the motor symptoms (such as depression or sleep disorder) provide a very limited scope for judgement. There are many more patterns , describing the relationship between the behaviour of the human body and Parkinson's. These patterns persist in the grey area, instead of just black or white, and are required to be learnt and studied to diagnose this disease rather precisely.

Thirdly, the traditional method does not only require expertise, but it also requires the time to put in manual labour for evaluating the patients based on their test and histories. As this disease is highly time-dependant, this process can rather prove detrimental to the patient's quality of life by the time the diagnosis is announced. Lastly, the lack of structure in the diagnosis across the medical industry has led to a haphazard criteria for diagnosis. Different sets of experts might believe in conducting different set of tests and examinations instead of a common consensus, hence leading to inconsistency and increased possibility of misdiagnosis.

This is where machine learning comes into the picture. Machine learning has made everything way more advanced as its tremendous capability to analyze bulk of data and provide an analysis has affected the medical industry greatly. Constraints in traditional methods can lead to wrong diagnoses, reversing efforts for awareness. It is for the same reason that a precise model with low false positives and false negatives is crucial for effective detection and diagnosis.

To address the restrictions that are posed by the traditional method, we have proposed a model that blends the concepts of ensemble learning with that of neural networks. By minimizing the probability of overfitting and improvising the generalization, this paper will be able to zero down to a single highly precise model that can be used to increase scope, curb subjectivity, create standardization and remove bias of diagnosis so that people can be saved from this life-impairing disorder to lead a healthy life.

II. Related Work

Multiple researchers have researched and tried to find the methodology that can be the best suited for detecting Parkinson's in individuals. Initially, the study began in early 2010s when Zuo et al. (2013) [1], commenced this research by applying a swarm optimization enhanced K-Nearest Neighbor technique using fuzzy logic to detect Parkinson's. This paper paved a way for further research for the subsequent publication of a paper by Bind et al. (2015) [2] that provided a comparative analysis of supervised machine learning techniques showing how Parkinson's detection accuracy touched new heights with each model application.

Over the course of the next two years, delving deeper into this subject, Dinesh et al. (2017) [3] implemented the boosting ensemble technique to adapt complex learning patterns and detect Parkinson's, with lesser complications in a viable manner. Following this, in the same year, Fayyazifar et al. (2017) [4], elevated the ensemble approach using the analogous voice features, by presenting a competitive review between two models using different ensemble techniques. This paper was a breakthrough as it changed the conventional deployment of bagging techniques for detection of medical problems.

It was astonishing to note that ensemble learning did increase the efficiency of detection, however, recently, Alzubaidi et al. (2021) [5] proved how neural network holds an integral role in combatting Parkinson's through detection due to the model's behaviour of analyzing non-linear relations from large chunks of data, which led to the birth of a more practical predictive detection mechanism. In concurrence with this discovery, Velmurugan et al. (2022) [6] created a state-of-the-art prototype using neural network and stacking classifier by combining both the efficient approaches, which proved to be the most practical and versatile one yet.

Upon reading the previous study conducted in this field, it was clear that ensemble learning was leveraging the main strengths of individual supervised machine learning algorithms, leading to a more flawless detection. This error-free trait of the models explored earlier, posed as a compelling motivation to use ensemble in this project for the development of our solution.

III. Dataset

The Parkinson's Data Set is a collection of biomedical voice measurements from 31 individuals, including 23 people with Parkinson's disease (PD). The dataset is intended for use in discriminating between healthy individuals and those with Parkinson's Disease based on voice measures.

The data was compiled and published by Max Little, Patrick McSharry, Eric Hunter, and Lorraine Ramig in 2008, and the study was focused on the suitability of dysphonia measurements for telemonitoring of Parkinson's disease. For this project, the previously compiled dataset has been referenced from a website which is a machine learning repository containing many such diverse datasets. [7]

The dataset was present online in ASCII CSV format, with each row representing an individual voice recording, and each column containing a specific voice measure. The size of the dataset (195x24) was such that it consisted of 195 recordings in total with 24 columns, for each, to be exact. Based on the number of patients and the rows, there are about six recordings per individual (195/31).

From the 24 columns, the first and the last columns are the 'name' of the individual with the recording number and the health 'status' of the individual respectively. The first column serves as a method for identification of the records and on the other hand, the last column indicates whether the individual has Parkinson's Disease or is healthy. The 'name' column follows an ASCII naming convention, and the 'status' column follows a binary convention, with a value of 1 indicating that the individual has the disease and a value of 0 indicating that the individual is healthy.

The rest of the 22 features or voice measures have been extracted from the audio signals recorded during the experiment. They also follow a numerical convention. The extracted features are varied and comprehensive, covering fundamental frequency, variation in amplitude and fundamental frequency, ratio of noise to tonal components in the voice, nonlinear dynamical complexity measures, and signal fractal scaling exponent. Some of the specific measures include the average, maximum and minimum vocal fundamental frequency, several measures of variation in fundamental frequency, and of variation in amplitude.

The data also includes nonlinear dynamical complexity measures such as RPDE and D2, which are used to assess the complexity of the signals generated by the vocal cords. Finally, three nonlinear measures of fundamental frequency variation are included in the dataset: spread1, spread2, and PPE.

Exploratory Data Analysis

Upon performing basic preliminary data analysis on the dataset, two major dataset characteristics are studied. As shown in Fig. 1, the function 'missingno' clearly shows that there are no null values present in either the rows or columns. Hence, eliminating a chunk of the preprocessing required before the model could be trained.

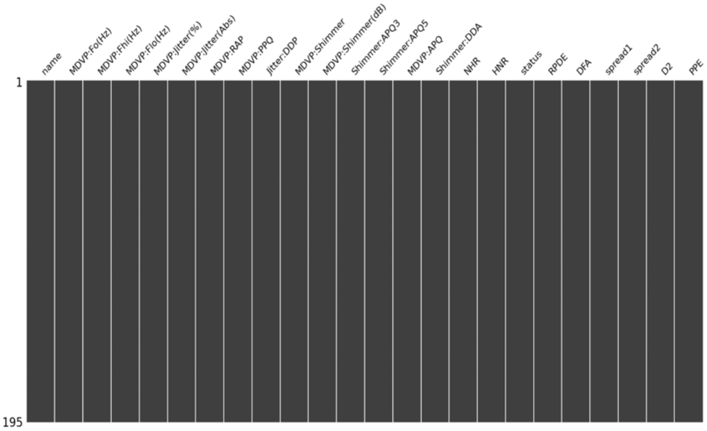

Fig. 1: Exploratory Data Analysis to examine the dataset for null values using 'missingno' function

The second characteristic examined is the distribution of the target class using 'value_count' function on the 'status' column. The function helps in providing a clear picture about how the 'status' data distribution is to be handled. As shown in Fig. 2, the classes are imbalanced to a great deal and would require some rectification to make sure there is no significant bias post the model training and hence no impact on the evaluation metrics.

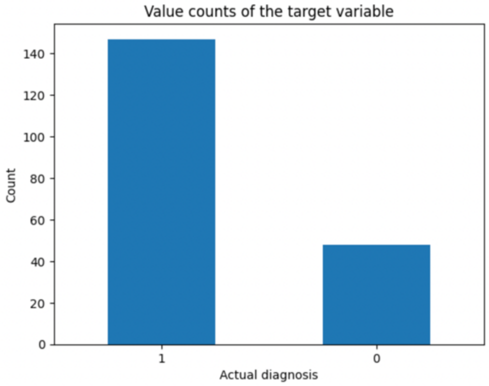

Fig. 2: Exploratory Data Analysis to examine the disturbtuioin of the 'status' (target) variable using 'value_count' function.

Overall, this dataset provides a valuable resource for researchers and clinicians seeking to develop and evaluate tools for the diagnosis and management of Parkinson's disease. The comprehensive voice measures included in the dataset offer insights into the vocal characteristics of individuals with PD, and may help to identify biomarkers that could be used to support earlier diagnosis as well as more effective management of the disease.

IV. Methodology

A. Data Preprocessing

In machine learning, data preprocessing is a crucial step that involves cleaning, transforming, and scaling the data to make it suitable for use in training a machine learning model. The quality of the data directly impacts the accuracy of the model, so data preprocessing is essential for producing reliable results.

For this dataset, a few simple steps had to be performed to make sure the data was in its best state before training the model on it. The first step in implementing a machine learning model using a dataset is to load the data into a Pandas data frame from a CSV file. However, it is crucial to ensure that the dataset is complete, clean, and error-free before analysis. This can be achieved by thoroughly examining the data for any missing values or inconsistencies that could affect the model's accuracy.

Once the dataset is deemed suitable for overview, data analysis is performed to gain a better understanding of its structure. Properties such as 'info', 'shape', 'head' and 'size' are used to explore the data and identify any issues that may impact the model's performance.

As seen earlier, no null values had to be eliminated, and so, in order to address the other preliminary requirements for training the dataset, a function termed as 'myPreProc function' is used to ready the data. The function drops the 'name' column, as it is not necessary for the training of our proposed models and returns the feature variables (X) and target variable (y).

B. Model Training

Moving forth, for the training process, the feature variables (X) and target variable (y) are separated. This is done by dropping the 'status' column to obtain X and selecting only the 'status' column to obtain 'y'. A 70-30 data split method is implemented , using the scikit-learn library's 'train_test_split' function, by splitting the data into training and testing sets with 70% (approximately 137 records) and 30% (approximately 58 records) respectively. A random seed of 42 is also specified which ensures that the same sequence of random numbers is generated, resulting in the same data split and consistent model performance.

However, the data is still not ready to be scaled yet. Due to the presence of the class imbalance, this problem had to be addressed. Synthetic Minority Over-sampling Technique has been used is to balance the imbalanced training data. It is applied on the training set using the SMOTE() function. The random state parameter is set to 42 for reproducibility.

The SMOTE function resamples the training set by creating synthetic samples of the minority class (in this case, the negative class) until the number of instances in both classes is equal. The resulting resampled training set is then used in the data scaler further, prior to training the machine learning model.

The standard scaler functionality is used to adjust the scale of the values in the training and testing data for easier analysis and comparison. This is done by transforming the data to have an average value of 0 and a typical size of 1. This in turn makes sure that all the different features in the data are being compared on a similar scale.

The main aim of this project is to implement ensemble machine learning techniques to compare and analyze the applied models based on the evaluation metrics that are recall, f1-score, precision, and accuracy. To carry forward the procedure, the dataset, post processing is trained on seven different ensemble models to be exact. Three models implement bagging, three implementing boosting and one model implements the stacking technique.

The three ensemble techniques that have been implemented on this data are- bagging, boosting and stacking. For bagging, the three models are created by using Random Forest, Decision Tree and K-Nearest Neighbour as base estimators for bagging classifiers. Similarly, for boosting, the three models use Adaboost, Gradient Boosting and XGBoost classifiers respectively. Lastly, for stacking the base classifiers were Support Vector, Gaussian Naïve Bayes and K-Nearest Neighbors Classifier. The output of these base classifiers is used as input to a logistic regression meta classifier.

Once, the models are trained as per requirement, the best model is chosen out of the lot. To assess the quality of the predictions of the chosen model, a 'Receiver Operating Characteristic - Area Under the Curve' (also known as ROC-AUC) graph is plotted. This enables one to understand how the true positive instances are distinguished from the false positives which is a highly essential characteristic for the determination of Parkinson’s as it directly correlates to missed diagnosis and misdiagnosis respectively.

Upon brief analysis of the ROC-AUC graph, the model is optimized to produce better results. Using the GridSearchCV functionality, optimal hyperparameters are identified for the picked model. The then trained model is tuned based on the recognized parameters to yield the highest performance metrics for the model.

V. Result & Analysis

A. Evaluation Procedure

The evaluation procedure involves emphasis on the four-evaluation metrics due to the classification nature of the problem. These four metrics, accuracy, precision, f1-score and recall, have been calculated for each performance using confusion matrices. The results are stored in a pandas data frame and printed as a grid using the tabulate() function.

Post calculation, a horizontal bar chart, displaying the performance of each of the models, has been plotted to show the score metrics for each model as bars of different colors. The models are first compared within the same type of classifier such as Bagging and Boosting and then towards the end, all the models have been plotted on a single chart, as shown in Fig. 3, for a clear view.

Since sensitivity plays a big role in reducing the number of false negative instances, and precision accounts for the minimization of false positive instances, special consideration has been given to recall and precision, respectively, during the evaluation procedure for the detection of Parkinson’s.

B. Preliminiary Analysis

When glanced at the final resultant graph as shown in Figure 3, it is evident that the performances of all models range between 88% to 97% range in the evaluation metric scores. Also, the stacking classifier outperforms each model and projects the highest bar from all the rest. With XGBoost not far behind, it gives a fair and tough competition to the stacking classifier.

However, it is essential to note that the Adaboost classifier, performs the worst out of the lot, taking the last position as the on the graph. Although the initial preliminary reading of the bar graph is not enough to understand the in-depth analysis of model performance, it does show how largely the results vary between the models themselves.

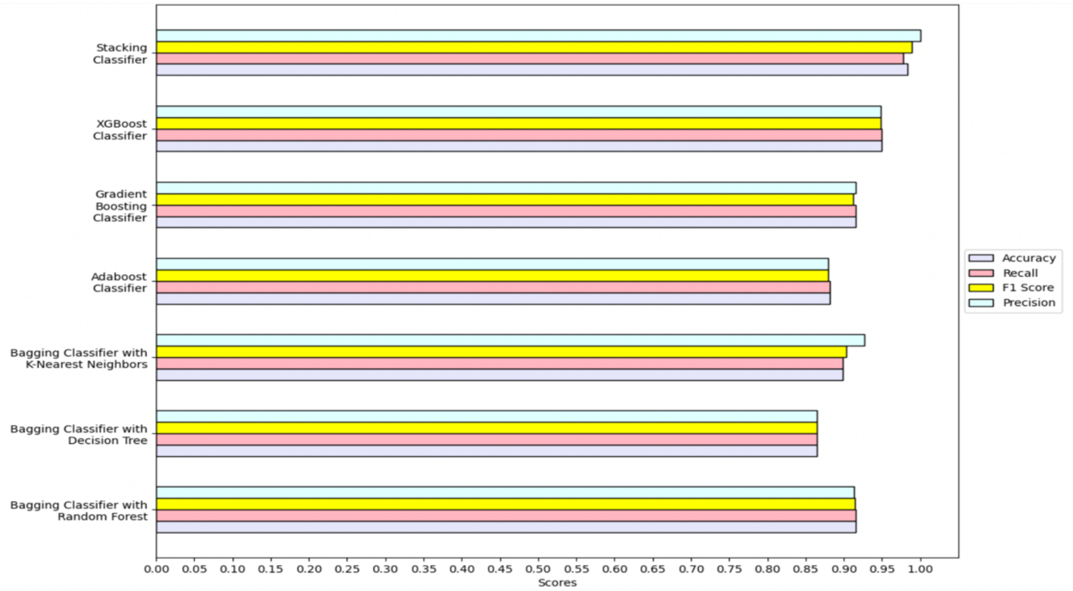

Fig. 3: A graphical representation comparing all the models to evaluate the performance of each ensemble technique on unseen data

C. Initial Results

TABLE I: Evaluation Metric Values For All Ensemble Models

Model	*Accuracy*	*Recall*	*F1 Score*	*Precision*	*Confusion Matrix*
Bagging Random Forest	0.9152	0. 9152	0.9142	0.9139	[[12 3] [2 42]]
Bagging Decision Tree	0.8644	0.8644	0.8644	0.8644	[[11 4] [4 40]]
Bagging KNN	0.8983	0.8983	0.9030	0.9273	[[15 0] [6 38]]
Ada Boost	0.8813	0.8813	0.8799	0.8792	[[11 4] [3 41]]
Gradient Boosting	0.9152	0.9152	0.9119	0.9153	[[11 4] [1 43]]
XGBoost	0.9491	0.9491	0.9485	0.9486	[[13 2] [1 43]]
Stacking Classifier	0.9661	0.9772	0.9772	0.9772	[[14 1] [1 43]]

Referring to the Table I above, at first glance at the bagging classifiers, it is evident that the three models perform somewhat similarly in terms of precision, all ranging between 86%- 92%

indicating that the false positive rate lies low. Based on the confusion matrix, the K-Nearest Neighbour classifier tended to have the lowest rate of false positive detection as compared to the other two bagging classifiers. And on the other hand, the Decision Tree model proved to be less effective in terms detecting precisely.

Keeping in mind the type of diagnosis this dataset offers, it is important to note that alongside precision, recall played a big role and surprisingly, in this field, Random Forest excelled immensely, which was quite a contrast to the behaviour Decision Tree exhibited.

The recall and precision directly affect the f1-scores to make sure the model performs both, the actions of correctly depicting positive instances while keeping false positivity rate low, in harmony with each other as a combination. In terms of f1-score, amidst all the bagging classifiers, the Random Forest model performed the best and K-Nearest Neighbour , the worst. Lastly, when accuracy was taken into consideration, the bagging classifiers with Random Forest, Decision Tree, and K-Nearest Neighbors achieved high accuracy scores, indicating that they can correctly classify most instances. But as usual, from all of them, the Random Forest-based model performs the best showing the highest accuracy.

Even though K-Nearest Neighbour showed tremendous precision, all the trends cannot be discarded due to one value in the overall metric analysis and so Random Forest was deemed as the model that showed the best performance among bagging classifiers

Moving on to the boosting classifiers, we can see that all three models - Adaboost, Gradient Boosting, and XGBoost - achieved high accuracy scores, with XGBoost performing the best. The confusion matrix also indicates that all three models had a low false positive rate, with XGBoost achieving the lowest rate yet. In terms of recall, all three models achieved high scores, indicating that they were able to correctly identify most of the positive instances. However, Adaboost had a slightly lower recall score as compared to the other two models. Precision scores were also high for all three models, indicating that false positive rate was low. Much like earlier, XGBoost had the highest precision score, followed by Gradient Boosting and Adaboost fell at the last position. The f1-score is a combination of both precision and recall, and XGBoost once again had the highest f1-score amongst all the three boosting classifiers.

In conclusion, the boosting classifiers - Adaboost, Gradient Boosting, and XGBoost - all performed well in terms of accuracy, precision, recall, and f1-score, with XGBoost emerging as the best performing model in most of the evaluation metrics.

Lastly, the seventh or the most outstanding model, the Stacking classifier broke records for all the four metrics and gained an exceptionally high precision score, almost nearing 100%. The number of incorrect identifications were also noted to be the least as per the confusion matrix the stacking classifier portrayed. Overall, the stacking classifier performed the best out of all the seven models with XGBoost as the runner up and Bagging with Random Forest following closely behind.

D. ROC-AUC Analysis

Upon plotting the ROC-AUC curve for the best classifier, that is the stacking classifier in this case, the Fig. 4 was obtained.

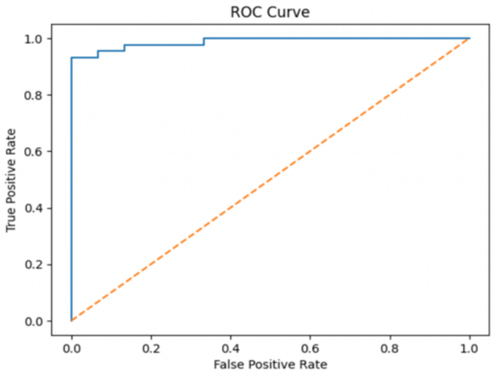

Fig. 4: ROC-AUC graph to visualize the quality of predicitions made by the stacking classifier on the test data.

As seen in Fig. 4, the stacking classifier, even though outperformed all the other models, but the linearity of the model failed to learn the complex patterns that the relationship between the feature and target variable posed. Due to the nonlinear nature of the training data, a huge chunk of information was being overlooked due to the linear nature of the meta model of the classifier. This led to a chaotic mess in the distinction between the true positive and the false positive rate in the graph due to which the quality of prediction is impacted.

VI. Proposed Remediation

Due to the presence of drawbacks in the stacking classifier as seen earlier, a remediated model is proposed which replaces the older logistic regression meta model with a neural network to study the complex pattern that influences the target variable indirectly. As seen in Fig. 5, the meta model is a single layer perceptron consisting of three layers in total, the input, hidden and the output layer. The newer model has been derived by drawing inspiration from an existing literature article [6] , that showcases how neural networks learn the intricate multidimensional trend between features and target.

The model uses single layer perceptron due to a few reasons, mainly - Firstly, as the dataset does not have the size and diversity it is almost impossible to implement deep learning models such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) on the training data which can even include irrelevant details and noise, and hence perform poorly on the unseen data, leading to overfitting. Secondly, the time taken by a single layer perceptron is always lesser than a multi-layer classifier due to the number of layers.

VII. Optimization

As the final classifier got modified to rectify the drawbacks and incorporate the best practices, prior to evaluating the new model on the test data, the optimal hyper parameters were identified for the new stacking classifier using the GridSearchCV.

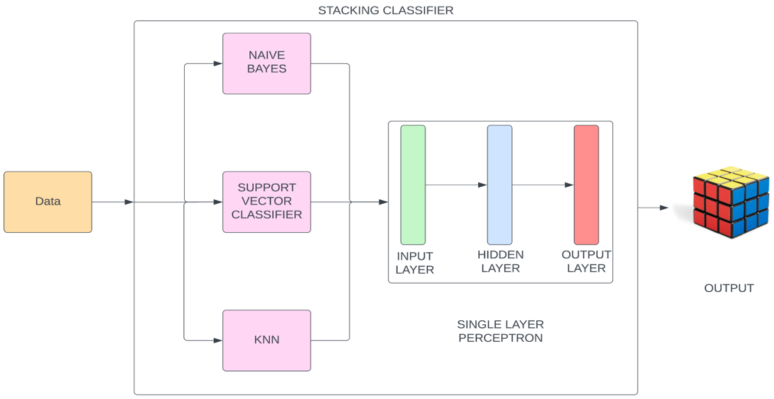

Fig. 5: Architecture of the newly proposed stacking classifier with single layer perceptron (neural network) as the meta model

The hyperparameters that were used during this search alongside the values that were optimal for each are shown in Table 2 below.

TABLE II: Hyperparameter Analysis Based On The Gridsearchcv

Name of hyperparameter	***Hyperparameter model***	***Optimal value***	***Purpose of choosing hyperparameter***
Number of neurons	Single Layer Perceptron(meta)	16	To identify the patients that show nonlinear patterns of symptoms.
Regularization parameter	Support Vector Machine (base)	0.1	To identify the ill patients from the healthy ones.
Kernel Coefficient	Support Vector Machine (base)	1	To identify the patients that show unusual symptoms.
Number of neighbors	K-Nearest Neighbors (base)	2	To identify the patients with a range in the severity of their symptoms.

VIII. Final Outcome

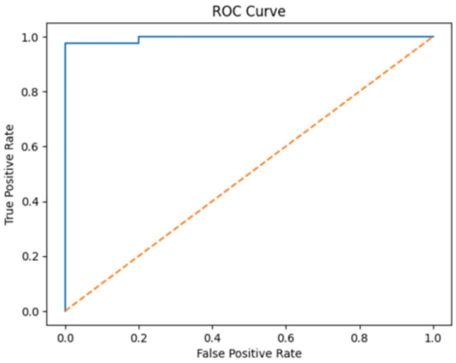

Fig. 6: ROC-AUC graph to visualize the quality of predicitions made by the tuned new stacking classifier (with neural network) on the test data.

The graph in Fig. 6 that has been obtained, post creation of the new stacking classifier, when tuned, produces an extremely smooth curve. The performance of the stacking classifier at a 98% accuracy was immensely exceptional. The classifier distinguished almost all the false positive from the true positives correctly. The neural network was able to avoid all misdiagnosis and had only one missed diagnosis in the detection on unseen data. The precision tended towards 100% which was a bit exemplary, but keeping in mind how small the dataset was, the model learnt all the patterns elaborately from the training data and predicted well on the test set.

IX. Conclusion

The study this project conducted shows how when integrated with neural network, the stacking ensemble technique can be used as a real-life machine learning model to classify patients suffering from this disease. In future, this model can be trained on a larger dataset including different features (such as symptoms other than vocal biomarkers found in patients over time), to test the relevance of the stacking classifier and create a robust model.

References

[1] W.-L. Zuo, Z.-Y. Wang, T. Liu, and H.-L. Chen, “Effective detection of Parkinson’s disease using an adaptive fuzzy k-nearest neighbor approach,” Biomedical Signal Processing and Control, vol. 8, no. 4, pp. 364–373, Jul. 2013, doi: https://doi.org/10.1016/j.bspc.2013.02.006. J.

[2] S. Bind, A. Tiwari, and A. Sahani, "A survey of machine learning-based approaches for Parkinson disease prediction," Int. J. Inf. Technol. Comput. Sci., vol. 6, pp. 1648-1655, 2015.

[3] A. Dinesh and J. He, "Using machine learning to diagnose Parkinson's disease from voice recordings," *2017 IEEE MIT Undergraduate Research Technology Conference (URTC)*, Cambridge, MA, USA, 2017, pp. 1-4, doi: 10.1109/URTC.2017.8284216.

[4] N. Fayyazifar and N. Samadiani, "Parkinson's disease detection using ensemble techniques and genetic algorithm," *2017 Artificial Intelligence and Signal Processing Conference (AISP)*, Shiraz, Iran, 2017, pp. 162-165, doi: 10.1109/AISP.2017.8324074.

[5] M. S. Alzubaidi et al., “The Role of Neural Network for the Detection of Parkinson’s Disease: A Scoping Review,” Healthcare, vol. 9, no. 6, p. 740, Jun. 2021, doi: https://doi.org/10.3390/healthcare9060740.

[6] T. Velmurugan and J. Dhinakaran, “A Novel Ensemble Stacking Learning Algorithm for Parkinson’s Disease Prediction,” Mathematical Problems in Engineering, vol. 2022, pp. 1–10, Jul. 2022, doi: https://doi.org/10.1155/2022/9209656.

[7] https://archive.ics.uci.edu/ml/datasets/parkinsons

[8] Github- https://github.com/kaulakh4/Machine-Learning-Final-Code-Group-11.git

CONTRIBUTION:

○ Kanika Gandhi : Implementation and Comparison of Ensemble Models, Model Result Analysis , Hyperparameter Optimization.

○ Karanpratap Singh Aulakh: Related Work, Data Preprocessing, New Model Implementation (Stacking +NN), Final Analysis.

○ Kiranjot Kaur: Study of Need Analysis and and Drawbacks of the Problem, DataSet Analysis, Model Training (Ensemble Technique).