Ensemble Learning for Parkinson's Disease

Aaparna Balan - S1042839
Puja Prakash - S1039329

I. INTRODUCTION

Parkinson's disease (PD) is a degenerative neurological disorder that affects the body movement, including the presence of tremors and stiffness by reducing the dopamine levels in the brain [1]. The disease can show its effects in a speech in many different ways. The person affected by PD tends to speak in low volume and in one tone (hypophonia), difficulty in articulating sound (dysarthria), and reduced pitch type (monotone). It can also cause emotional changes and decreases the ability of thinking and to perform day to day activity (dementia).

For the diagnosis, the clinician takes the patient's neurological history and analyses the motor skills in different scenarios [1]. Due to the lack of laboratory tests, the diagnosis becomes difficult especially in the early stages which has no severe motor effects. We propose machine learning techniques to make the screening process easier and more convenient which does not require a clinical visit. As the voice recordings exhibit characteristic PD patients these features can be analyses for the diagnosis which can result in an effective screening approach before meeting the clinician. This saves a lot of time and makes the process quicker.

We used Parkinson's dataset from Kaggle and performed ensemble learning which combines the results of different algorithms to one predictive model. It is required to take the right decision during the screening and the tonality of each person varies. To improve the performance and to make appropriate decisions, combining various model predictions can reduce bias making the analyses more accurate.

A. *HYPOTHESIS*

The speech recordings help us to detect the emotional changes of the speakers by extracting the important feature and using it in the model. The main aim of choosing ensemble learning is to show how good the combined results can make the screening process simpler.

- Does the ensemble model perform better than the other machine learning algorithms to classify patients with or without Parkinson's disease taking into account the patients' variability and complexity of the disease?

Considering these factors it is convenient to use ensemble learning as it is the combination of multiple base models and advances in the decisions while analyzing the patients' recordings and segregating them with or without the disease [2]. In this paper, we study whether ensembles of decision trees can have better results as compared to other algorithms for the prediction of Parkinson's disease.

II. RELATED WORKS

In most of the cases, biomedical disease diagnosis problems are likely to have two class predictions. The goal of these problems is to map data samples into one of the groups (i.e. benign or malignant) with good accuracy. There are a lot of related publications on ensemble learning which shows a good improvement in performance by combining the results of different classifiers [3]. In the paper of Little et al. [4], dysphonia measures are used to evaluate the performance of the algorithms in the diagnosis of PD. Neural network comparison was performed by Das, Resul. (2010) [5], which was also a fascinating approach and the results were more or less similar to other classifiers. The paper [6] also gives an understanding of the necessity of initial screening and theoretically explains the effects of the disease. This allowed us to understand the need for producing a model that can highly benefit in classifying the affected patients which can help in faster treatment.

On the other hand, feature ranking is useful to attain knowledge of data and to deduce good features. The article [7] explores the performance of combining linear support vector machines with various feature ranking methods. The paper [8] also shows that reviews ensemble-based systems may be more beneficial than single classifier and various procedures through which the individual classifiers can be combined. It also illustrates the most popular ensemble-based algorithms, such as bagging, boosting, AdaBoost, stacked generalization, combination of rules, and decision templates. We used this as a reference to pick the classifiers.

III. DATASET

The project uses Kaggle dataset [1] containing voice recordings of patients from which features such as different measures of variation in fundamental frequency (maximum, average, etc), amplitude (Shimmer), measures of the ratio of noise to tonal components in the voice (NHR, HNR), nonlinear dynamical complexity (RPDE, D2), Signal fractal scaling exponent and nonlinear measures of fundamental frequency variation (spread1,spread2, PPE) were extracted to build machine learning algorithms. The target variable was taken as the health status of the patients i.e 0 for healthy patients, 1 for patients suffering from Parkinson disease.
The data set consists of 195 recordings and 24 features. More than 75% of the data have Parkinson's disease than people not having Parkinson's.

IV. DATA EXPLORATION

Uni-variate analysis such as frequency plots were created to check the distribution of features. Fig 1 shows the distribution of various measures of the vocal fundamental frequency (Minimum, maximum, and average). The feature average vocal fundamental frequency is normally distributed whereas it's minimum and maximum values are positively skewed indicating a higher number of values being observed at lower frequencies.

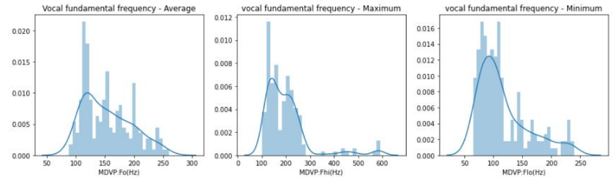


Fig. 1. Frequency distribution

A similar analysis was conducted for all the features and the results showed that most of the features are positively skewed.

V. PRE-PROCESSING

As the features need to be uncorrelated for training the model, correlation plot was observed as shown in Fig 3. It is seen from the plot that MDVP Jitter % is highly correlated with Jitter (Abs), MDVP RAP, MDVP PPQ, NHR variables, HNR has high correlation with variables MDVP Shimmer (dB), Shimmer APQ3, Shimmer APQ5, MDVP APQ, Shimmer DDA and MDVP Shimmer is correlated to MDVP Shimmer (dB), Shimmer APQ3, Shimmer APQ5, MDVP APQ, Shimmer DDA features. One of the reason for this is because of the features being derived from each other.

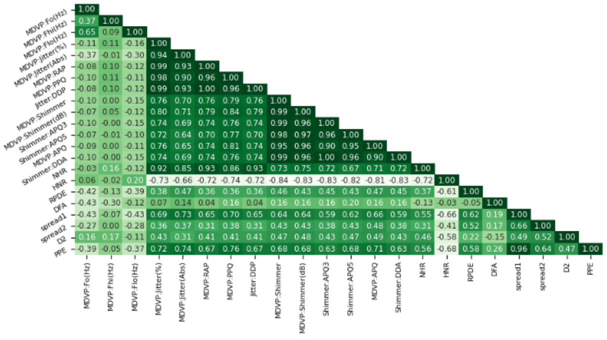


Fig. 2. Correlation plot

The correlation between the target variable Status and other independent variables was also observed. Fig 4 Shows the distribution of minimum and maximum vocal fundamental frequency for status 0 and status 1. People with minimum vocal frequency of 250Hz and above are less likely to suffer from Parkinson's disease whereas for maximum vocal fundamental frequency patients with values ranging between 100-200 Hz show higher evidence of having Parkinson's disease.

Based on the correlation values in fig 1, the features with high correlation of 80% and higher (MDVP:Shimmer , MDVP:Jitter(%) , HNR) was removed from the training set to avoid misinterpretation of training results.
As most of the features had varied units of measurement, the traiind data set was normalized.

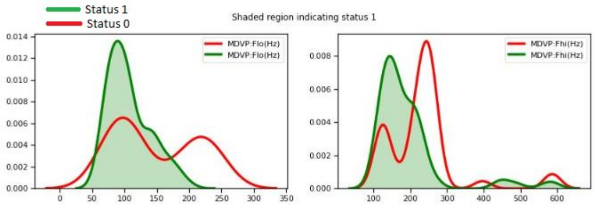


Fig. 3. Frequency distribution

VI. SETUP

A. METHODOLOGY

To build the model we used 80 % as train dataset and the remaining to test the model. The variance has been checked to determine the features which affects the target variables. We also did feature scaling as their exists different units in the features [7]. As ensemble learning combines the results of different classifiers to improve variance, bias or prediction stacking. Classification helps us to determine the categories to which the new observations can be included with the use of classifiers. Here we use five different classification algorithms which is discussed as follows.

B. LOGISTIC REGRESSION

A sigmoid function which is similar to logistic function is used for which the output of the classification is between 0 and 1. Logistic regression determines the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables [10]. As unlike linear regression the outcome is limited to few possible values and helps us to categorize the Parkinson's disease affected patients.

C. KNN

The idea of similarity can be captured in the K nearest neighbor algorithm (KNN) for which the Euclidean distance is calculated from all points and classifies each data point. The existing data allows us to estimate the distance of new data in the sample data set and is classified based on the k neighborhood. To choose the optimal 'k' value we scaled the values from 3 to 35 and used misclassification error to choose the optimal number of neighbors which is 25.

D. SVM

SVM is a technique that transforms the data and determines the optimal boundary between the possible outputs. To determine the optimal boundary we used hyperparameters - gamma and C. It is evident that with low gamma value the influence of the training reaches far and provides low bias. On the other hand, C is the cost of misclassification for which a large c value gives low bias and high variance and vice versa. We chose the optimal values for gamma and C as 0.05 and 3, after trying out with different values.

E. STACKING

We used different classifiers to predict Parkinson's disease, for which we used a novel technique called stacking which allows us to combine different weak learners by training the model to produce results based on multiple predictions. To do this, a learner (L) (i.e.) base models are used and the meta-model that combines them. The stacking model architecture used involves 2 or more base models which are also known as level-0 and level-1 models.

- Level-0 Models: Models fit on the training data and whose predictions are combined.
- Level-1 Model: Model that learns how to best combine the predictions of the base models [9].

Here we used Logistic regression, SVM, and KNN as our base models, and their predictions are compiled and the support vector classifier (SVC) is used to fit the data which is provided by the base models to provide the best fit. The gamma and C values for SVC are 0.07 and 4 which produced improved accuracy and discussed in section VII. The stack model used the base models as the estimators which have to be stacked together, and level-1 as the final estimator which combines the base estimators and the cv as 5 that is the cross-validation splitting strategy used to train final estimator (k folds).

F. RANDOM FOREST

It is an ensemble learning classifier which generates decision trees randomly reducing over-fitting of the model by averaging the results. Here we used the number of trees as 100 and for the criterion we used entropy. We used this metric to measure the information gain and measures on how to reduce the uncertainty for the label and 8 features were used while looking for the best split at the node.

G. ADAPTIVE BOOSTING

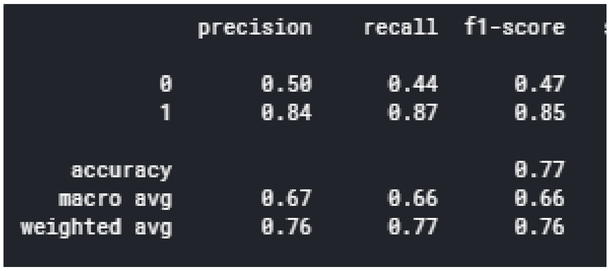
The ada-boosting is an ensemble technique that is used to fit the weak classifiers and improves the performance for each iteration. The number of estimators is used to terminate the boosting and we used100 estimators for our model.

VII. RESULTS

Ensemble methods such as Random forest and Adaptive boosting were compared with other machine learning algorithms such as SVM, KNN, and logistic regression.

A. Logistic Regression

The logistic regression model gives an accuracy of 77%. A ratio in target variable for 1s (Indicating patient having Parkinson's disease) and 0s (Indicating patients not having Parkinson's disease) is 75% to 25%, F1 score can be taken as 0.85, recall as 0.87 and precision as 0.84.

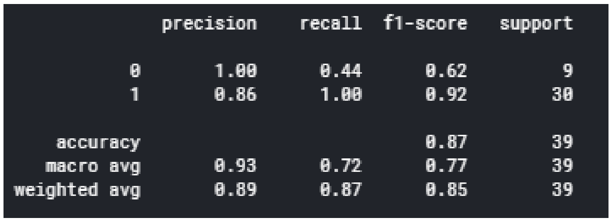


```
              precision    recall  f1-score

           0       0.50      0.44      0.47
           1       0.84      0.87      0.85

    accuracy                           0.77
   macro avg       0.67      0.66      0.66
weighted avg       0.76      0.77      0.76
```

Fig. 4. Logistic Regression model accuracy

B. K nearest neighbors

KNN gives an accuracy of 87% by considering 25 nearest neighbors and euclidean distance as the distance metric. The model results in an F1 score of 0.92, recall as 1, and precision as 0.86.

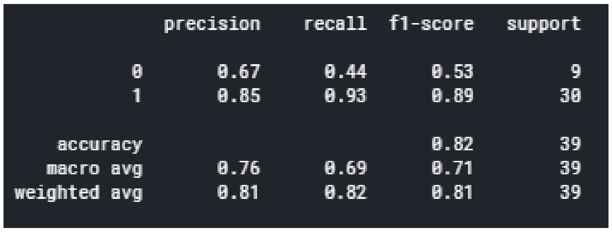


```
              precision    recall  f1-score   support

           0       1.00      0.44      0.62         9
           1       0.86      1.00      0.92        30

    accuracy                           0.87        39
   macro avg       0.93      0.72      0.77        39
weighted avg       0.89      0.87      0.85        39
```

Fig. 5. KNN model accuracy

C. Support Vector Machines

SVM gives an accuracy of 82% when trained with hyperparameter values as gamma = 0.05 and Cost as 3.As training data has 75% to 25% ratio for patients suffering from Parkinson's to healthy patients, F1 score can be considered as 0.89, precision as 0.85 and recall as 0.93.



```
              precision    recall  f1-score   support

           0       0.67      0.44      0.53         9
           1       0.85      0.93      0.89        30

    accuracy                           0.82        39
   macro avg       0.76      0.69      0.71        39
weighted avg       0.81      0.82      0.81        39
```

Fig. 6. SVM model accuracy

D. Model Comparison

To prevent overfitting cross-validation method was used where the training data was split into 15 k-folds and accuracy was compared between the 3 models - SVM, Logistic, and KNN. Fig 7 shows the comparison between these models. The mean F1 score for SVM was 0.93, for KNN 0.91 and Logistic 0.90. This shows that SVM performs better than other algorithms.

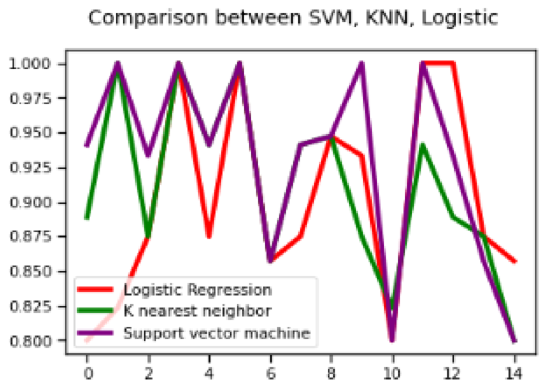
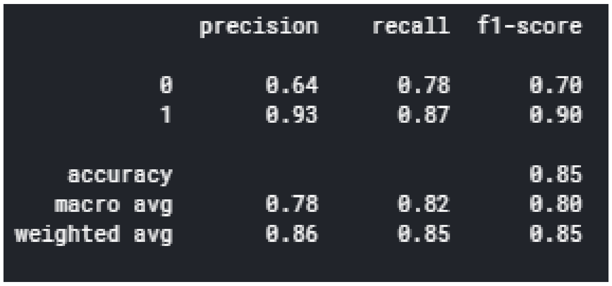


Fig. 7. Comparison between SVM, KNN and Logistic regression

E. *Ensemble Methods*

1) Stacking: Stacking was used as an ensemble learning method to combine the predictions from the classification models SVM, KNN, and Random forest using the meta-model SVC. The stacking classifier performed better than the individual classifiers as shown in Fig 8.



```
              precision    recall  f1-score

           0       0.64      0.78      0.70
           1       0.93      0.87      0.90

    accuracy                           0.85
   macro avg       0.78      0.82      0.80
weighted avg       0.86      0.85      0.85
```

Fig. 8. Accuracy using stacking classifier

To check how well the model classifies the patients with and without Parkinson's disease, ROC curver and AUC value was analyzed as shown in Fig 9

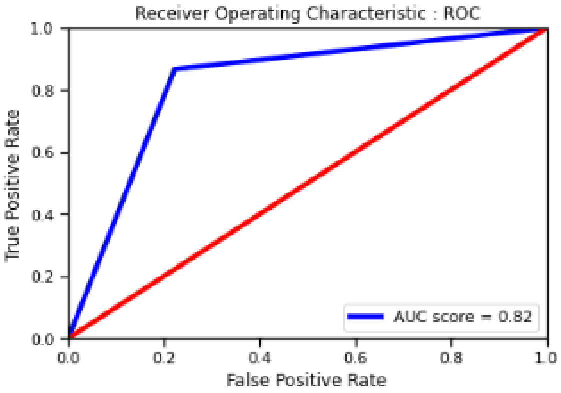
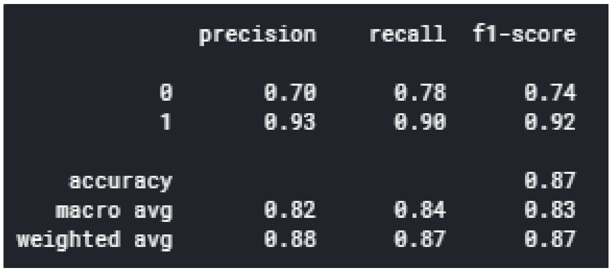


Fig. 9. ROC and AUC for stacking classifier

As the AUC score was 0.82, it can be concluded that the stacking classifier performs well in distinguishing between patients suffering from Parkinson's and healthy patients.

On a comparison between SVM, KNN, and Random forest it can be seen that support vector machines performs the best giving an accuracy of 82%. By using the Stacking model, the accuracy further increases to 85%, indicating that the ensemble technique improves model performance.

2) Random Forest: Random forest another ensemble model resulted in an accuracy of 85%, an F1 score of 0.90, a recall of 0.87, and a precision of 0.93 on combining the prediction results from 100 decision trees.



```
              precision    recall  f1-score

           0       0.70      0.78      0.74
           1       0.93      0.90      0.92

    accuracy                           0.87
   macro avg       0.82      0.84      0.83
weighted avg       0.88      0.87      0.87
```

Fig. 10. Random Forest Accuracy

ROC curve is shown in Fig 11, the AUC score was 0.84 indicating that the model performs better in distinguishing between the diseased and healthy patients than the Stacking classifier.

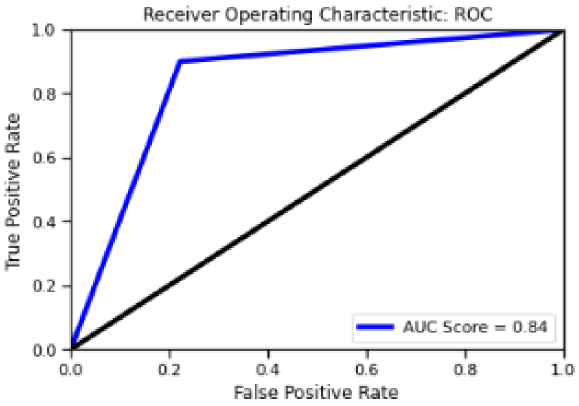


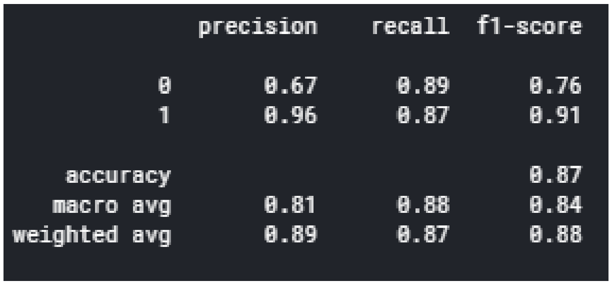
Fig. 11. ROC and AUC for Random forest classifier

3) Adaptive boosting: The adaptive boosting model resulted in an accuracy of 87% on considering 80 decision trees in sequence. F1 score obtained was 0.91, recall of 0.87, and precision of 0.96 on combining the prediction results from 80 decision trees in order.

ROC curve is shown in Fig 12, AUC score was 0.88 indicating that the model performs better in distinguishing between healthy and unhealthy patients than random forest and Stacking classifiers.

F. *Comparison between Ensemble methods*

On comparing the 3 ensemble methods, it was observed that Adaptive boosting and random forest give an accuracy of 87% whereas stacking gives an accuracy of 85%. AUC value for Adaptive boosting was 0.88 where as for random



```
              precision    recall  f1-score

           0       0.67      0.89      0.76
           1       0.96      0.87      0.91

    accuracy                           0.87
   macro avg       0.81      0.88      0.84
weighted avg       0.89      0.87      0.88
```

Fig. 12. Adaptive boosting Accuracy

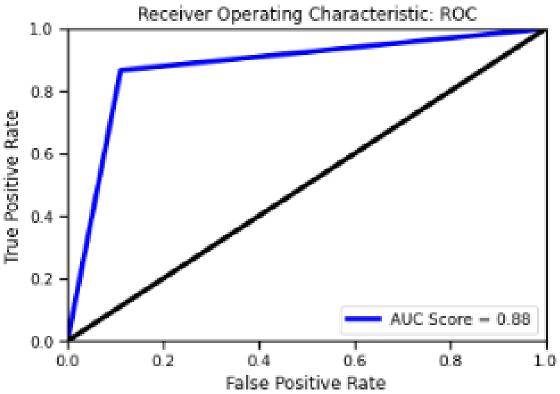


Fig. 13. ROC and AUC for Adaptive boosting classifier

forest it was 0.84, indicating that Adaptive boosting performs better than Random forest in terms of area under the curve.

VIII. DISCUSSION

We implement only the ensemble method for Parkinson's disease dataset and compared the results to that of other models. The performance of ensemble models was better than that of the other classifiers. We also combine the results of other models and compared the results by training random forest and Adaptive boosting which is ensembles model by itself. This helped us to understand how well ensemble techniques can improve the performance which allowed us to obtain an accuracy of 87% for adaptive boosting.

IX. AUTHOR CONTRIBUTION

Aaparna Balan

- Worked on Random Forest, Logistic Regression and SVM model
- Result Analysis
- Worked on report

Puja Prakash

- Worked on Adaptive boosting, KNN and SVM (included other hyperparameters)
- Result Analysis
- Worked on report

X. CONCLUSION

Parkinson's data set used speech information of the patients which is converted to vocal frequencies and is inputted to the classifiers to predict the healthy and Parkinson's affected patients. Here we used base models and ensemble classifiers and compared the results of both. It is evident that ensemble learning helps to improve more accuracy than that of the base models. On analyzing the results it was seen that Stacking on combining the predictions of 3 machine learning algorithms SVM, KNN, and Logistic performed better than individual classifiers indicating that ensemble technique improves model predictions. The stacking classifier was further compared with other ensemble techniques such as Random forest and Adaptive boosting. Adaptive boosting performed better than the other 2 ensemble models in terms of AUC but gave similar accuracy of 87% as that of random forest.

REFERENCES

[1] Kaggle, "Parkinson Data", retrieved from https://www.kaggle.com/malikasif123/parkinsondata

[2] Halawani S.M., Ahmad A. (2012) Ensemble Methods for Prediction of Parkinson Disease. In: Yin H., Costa J.A.F., Barreto G. (eds) Intelligent Data Engineering and Automated Learning - IDEAL 2012. IDEAL 2012. Lecture Notes in Computer Science, vol 7435. Springer, Berlin, Heidelberg

[3] Ozcift, A. SVM Feature Selection Based Rotation Forest Ensemble Classifiers to Improve Computer-Aided Diagnosis of Parkinson Disease. J Med Syst 36, 2141–2147 (2012), retrieved from https://doi.org/10.1007/s10916-011-9678-1

[4] Little, M., and McSharry, P., Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. Nature Precedings. 1-27, 2008

[5] Das, Resul. (2010). Das, R.: A comparison of multiple classification methods for diagnosis of Parkinson disease. Expert Systems with Applications 37, 1568-1572. Expert Systems with Applications. 37. 1568-1572. 10.1016/j.eswa.2009.06.040.

[6] Jyh-Gong Gabriel Hou and Eugene CLai. Non-motor Symptoms of Parkinson's Disease, retrieved from https://doi.org/10.1016/S1873-9598(08)70024-3

[7] Chang, Y., Feature ranking using linear SVM, JMLR: workshop and conference proceedings. 53-64, 2008.

[8] Polikar, Robi. (2006). Polikar, R.: Ensemble based systems in decision making. IEEE Circuit Syst. Mag. 6, 21-45. Circuits and Systems Magazine, IEEE. 6. 21 - 45. 10.1109/MCAS.2006.1688199.

[9] Stacking Ensemble Machine Learning With Python,retrieved from, https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/

[10] Statistical Solution, Logistic regression, retrieved from https://nl.wikipedia.org/wiki/Logistische$_{r}egressie$