How to Cite:
Renuka, D. K., & Rakshana, S. (2022). Prediction of Parkinson's disease using machine learning classification from voice analysis. *International Journal of Health Sciences*, *6*(S1), 11762–11771. https://doi.org/10.53730/ijhs.v6nS1.7866

Prediction of Parkinson's disease using machine learning classification from voice analysis

Dr. D. Krithika Renuka
Department of Computer Science (PG), PSGR Krishnammal College for Women
Email: krithikam@psgrkcw.ac.in

S. Rakshana
Department of Computer Science (PG), PSGR Krishnammal College for Women
Email: rakshana311299@gmail.com

***Abstract*---**"PARKINSON SYNDROME PREDICTION BASED ON AUDIO MEASURES" is to analyze and find the prediction efficiency that would be beneficial for the patients who are suffering from Parkinson and the percentage ratio will be reduced. Generally, in the first stage Parkinson can be cured by the proper treatment. So, it is important to identify the PD at the early stage for the betterment of the patients. The main purpose of this project work is to find the best prediction model i.e., the best machine learning technique which will distinguishes the Parkinson's patient from the healthy person. The techniques investigated are Random Forest, SVM, Logistic Regression and Adaptive Boosting.

Keywords*---Index Terms*—Parkinson, machine learning, prediction, accuracy.

Introduction

Parkinson's Disease (PD) is a degenerative neurological disorder marked by decreased dopamine levels in the brain. It manifests itself through a deterioration of movement, including the presence of tremors and stiffness. There is commonly a marked effect on speech, including dysarthria (difficulty articulating sounds), hypophonia (lowered volume), and monotone (reduced pitch range). Additionally, cognitive impairments and changes in mood can occur, and risk of dementia is increased. Traditional diagnosis of Parkinson's Disease involves a clinician taking a neurological history of the patient and observing motor skills in various situations. Since there is no definitive laboratory test to diagnose PD, diagnosis is often difficult, particularly in the early stages when motor effects are not yet severe. Monitoring progression of the disease over time

Manuscript submitted: 27 March 2022, Manuscript revised: 9 April 2022, Accepted for publication: 18 May 2022

requires repeated clinic visits by the patient. An effective screening process, particularly one that doesn’t require a clinic visit, would be beneficial. Since PD patients exhibit characteristic vocal features, voice recordings are a useful and non-invasive tool for diagnosis. If machine learning algorithms could be applied to a voice recording dataset to accurately diagnosis PD, this would be an effective screening step prior to an appointment with a clinician.

In this paper, we aimed to construct a predicting model for the Parkinson disease using modern techniques of machine learning, especially classification approaches to predict the status of health of the patients. The predictive model would help the clinicians to identify the disease and diagnose it at the right time. The clinical data is collected from the electronic medical records which is available at the UCI ML repository. The dataset is gathered by screening the voices of the patients. The Parkinson dataset consists of large range of voice measurements from 31 people, in which 23 patients affected from Parkinson disease where each column is a particular voice measure, and each row corresponds to one of 195 voice recordings from the column. The voice measurements have 23 unique attributes which assist in distinguishing the healthy patients from the one who are affected from Parkinson Disease. Machine learning algorithms such as Random Forest, SVM, Logistic Regression and Adaptive Boosting are compared to analyze which algorithm provides the highest accuracy of prediction for Parkinson's disease with good performance and fast implementation. The performances were evaluated with accuracy, sensitivity, specificity, positive predictive value and negative predictive value.

Related Work

Many researchers have developed Parkinson detection systems using machine learning techniques like as KNN, Naive Bayes, and SVM. Rahul. R [7] develops a technique known as computer-based analysis, which employs comprehensive empirical examination of Convolution neural networks (CNNs) to develop prognostic and diagnostic biomarkers for Parkinson's disease [4]. To predict Parkinson's disease, picture properties that are invariant to rotation, translation, and blurring are examined. [13] Genes, age, stroke, and diabetes were found to have a stronger link with the prediction's outcome. However, this research has limitations in multivariate feature analysis. With an accuracy of 74.93 percent, their suggested model was able to discriminate healthy people from Parkinson Disease patients.

Ioannis G.Tsoulus and Georgia Mitsi [8] successfully diagnosed patients with the condition using supervised classification methods such as deep neural networks. The machine learning models' peak accuracy of 75% outperforms non-experts' average clinical diagnosis accuracy (73.8%) and movement disorder specialists' average accuracy with pathological post-mortem examination as ground truth. The models used in this study produced useful results, although they fall short in terms of Parkinson disease detection and sensitivity rate.

Timothy Wroge and Yasin Serdar zkanca [9] present early evidence that artificial intelligence (AI) systems could allow anyone to distinguish Parkinson's disease (PD) patients from Healthy Volunteers (HV) who were matched for age and gender and served as controls. This research uses a variety of upper limb functional metrics to demonstrate it. PD patients were discriminated from healthy subjects using iMotor, a cloud-based software platform powered by artificial intelligence (HV). iMotor is a clinically certified automated device that conducts upper limb function testing using the sensing capabilities of a smart-screen tablet. Experiments were done utilising data gathered by iMotor following a two-year follow-up of Parkinson patients, and HV was discovered to automatically classify records into two categories: HV and PD.

Existing System

The existing method [5][6] implements the identification of Parkinson disease by the indicators like quality of speech, phonological and acoustic cues. Factors such as genes, age, stroke and diabetes [12] were found out to have higher correlation with the outcomes of the prediction. The handwritten notes from patients were also taken to diagnose Parkinson Disease [13]. This method uses Support vector machine [2] [1] (SVM), KNN, Stacking classification algorithms for prediction and the accuracy predicted to be 81.8%.

Proposed System

The project's design flow, as illustrated in Fig. 1, begins with the gathering and compilation of the dataset based on the criteria. The risk factors associated with the datasets are verified. After that, the datasets are pre-processed. When it is necessary to verify for random under sampling and without random under sampling, vibrations are classified. The datasets are examined in accordance with the risk progression assessment. The dataset's numerous properties were subjected to the machine learning technique. Goal is to classify the patients into the respective labels using the attributes from their voice recordings. A patient's Parkinson disease diagnosis has been predicted.

Module Description

The overall proposed system is classified into three modules.

- Collection of Clinical Data
- Data Pre-processing
- Model Building

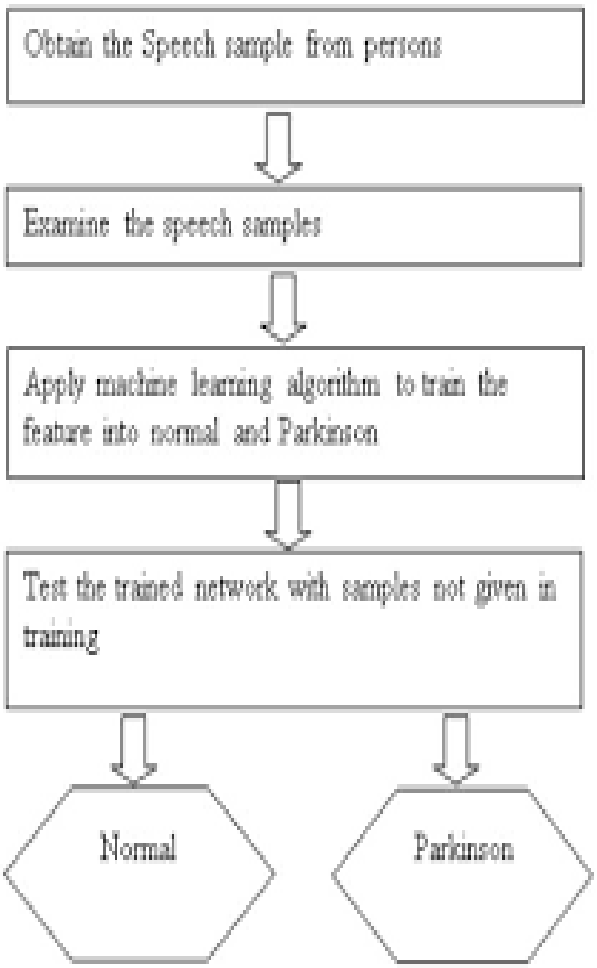

Fig. 1. Flow diagram of proposed systemCollection of clinical data

The Clinical Data is collected from the electronic medical records at UCI ML repository as shown in Fig. 2. All Parkinson affected patients were identified by the voice measurements [14]. Twenty-three predictor variables were collected from both healthy and Parkinson affected patients. These variables were MDVP, NHR, HNR, RPDE, D2, DFA. The classification models were used to identify Parkinson risk patients who should be facilitated with personalized medicine in the future.

```
a.dtypes

name                 object
MDVP:Fo(Hz)         float64
MDVP:Fhi(Hz)        float64
MDVP:Flo(Hz)        float64
MDVP:Jitter(%)      float64
MDVP:Jitter(Abs)    float64
MDVP:RAP            float64
MDVP:PPQ            float64
Jitter:DDP          float64
MDVP:Shimmer        float64
MDVP:Shimmer(dB)    float64
Shimmer:APQ3        float64
Shimmer:APQ5        float64
MDVP:APQ            float64
Shimmer:DDA         float64
NHR                 float64
HNR                 float64
status                int64
RPDE                float64
DFA                 float64
spread1             float64
spread2             float64
D2                  float64
PPE                 float64
dtype: object
```

Fig. 2. Clinical Dataset

Data-preprocessing

All variables with more than 50% missing value are removed as part of the pre-processing processes. To obtain a high-quality dataset, data imputation and standardization were also undertaken. To distinguish between normal and healthy changes in voice, Pitch Period Entropy (PPE), a dysphonia measure, is utilized to remove uncontrollable confounding effects such as noisy acoustic surroundings.

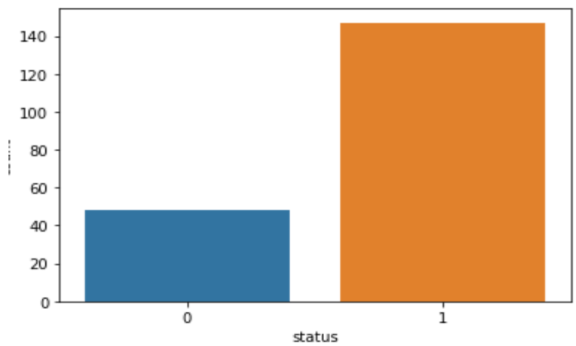

Fig. 3. Number of positive and negative detections of disease

```
[46] sns.boxplot(x='status',y='MDVP:Jitter(%)',data=Data)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0e1c2de550>
```

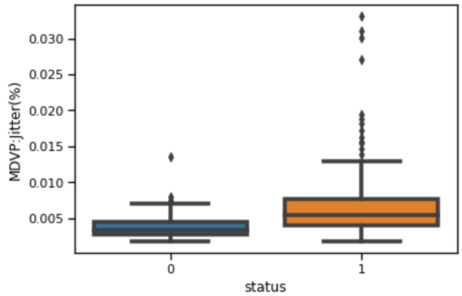

Fig. 4. MDVP values distribution in the dataset

Feature selection

Correlation coefficient features selection approach is used to indicate the relative importance of each feature when making a prediction. It differentiates the strong positive and negative correlation among features.

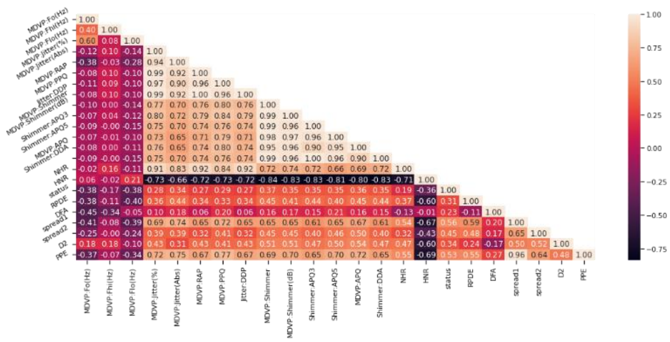

Fig. 5. Features correlation visualization using heat map

```
[77] sns.barplot(x=feature_imp, y=feature_imp.index)
     # Add labels to your graph
     plt.xlabel('Feature Importance Score')
     plt.ylabel('Features')
```

```
Text(0, 0.5, 'Features')
```

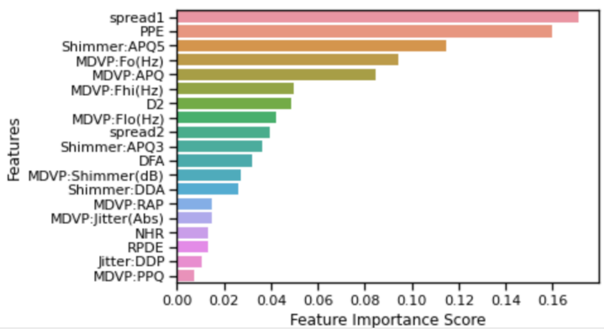

Fig. 6. Feature Selection

Model Building

Parkinson patients were accurately identified using predictive classifier algorithms. Logistic Regression, Random Forest (RF), SVM, KNN Stacking and Adaptive Boosting are the classification models for predicting Parkinson's disease.

Random Forest

Random Forest Algorithm performs better in terms of overall performance. Average Accuracy over data is 92% with Standard Deviation of 7% and F1-Score of 94% with Standard Deviation of 6%
The entire dataset is partitioned with similar amounts of data in each section, and a stratified k-fold cross validation procedure is used. The number of trees in the Random Forest determines the accuracy. Figure 6 depicts the classification report of random forest classifier predictions.

```
[73] print(classification_report(y_test, RandomForest_pred, digits=2))

              precision    recall  f1-score   support

           0       0.57      0.67      0.62        12
           1       0.91      0.87      0.89        47

    accuracy                           0.83        59
   macro avg       0.74      0.77      0.75        59
weighted avg       0.84      0.83      0.84        59
```

Fig. 7. Classification report of Random Forest classifier predictions

Logistic Regression, SVM, KNN

Logistic Regression has an accuracy of 79%, SVM has an accuracy of 84% and KNN has an accuracy of 64%. After comparing Logistic Regression, KNN, SVM algorithm we can conclude that SVM performed slightly better.

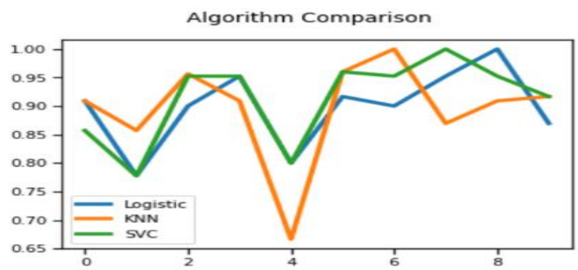

Fig. 8. Classification report of the algorithms

Adaptive Boosting

The weak learners in AdaBoost are decision trees with a single split, called decision stumps. → AdaBoost works by putting more weight on difficult to classify instances and less on those already handled well. → AdaBoost algorithms can be used for both classification and regression problem. This boosting algorithm takes as input a training set of N examples S= (x1,y1),......., (xm,ym)) where is an instance drawn from some space X and represented in some manner (typically, a vector of attribute values), and is the class label associated with . The set of possible labels Y is of finite cardinality k.

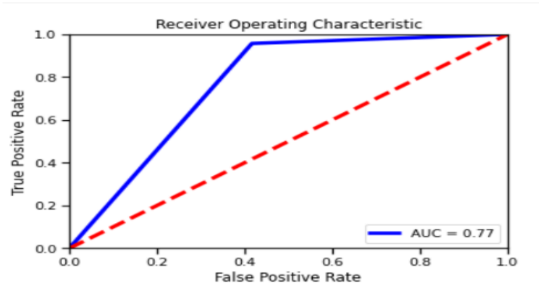

Fig. 9. AUC – ROC report of Adaptive Boosting classifier predictions

Results

Using ensemble classifiers like Random Forest and AdaBoostt, as well as a base classifier of Decision tree, the feature extracted data is further assessed to predict Parkinson disease. As shown in Fig. 10, the Random Forest algorithm model has the highest accuracy of 92.26 percent when compared to the other algorithm models like Random Forest and Adaboost. The accuracy, sensitivity, specificity, positive predictive value, and negative predictive value of the results were all assessed. Figure 10 depicts the final accuracy of each model.

```
Name = Logistic Regression , Mean Accuracy = 0.853846, SD Accuracy = 0.079441
Name = K-NN , Mean Accuracy = 0.839011, SD Accuracy = 0.110467
Name = SVM , Mean Accuracy = 0.876374, SD Accuracy = 0.084524
Name = Stacking , Mean Accuracy = 0.884066, SD Accuracy = 0.085878
Name = Random Forest , Mean Accuracy = 0.920330, SD Accuracy = 0.074501
Name = Adaptive Boosting , Mean Accuracy = 0.890659, SD Accuracy = 0.065955
```

Fig. 10. Final Accuracy of each model

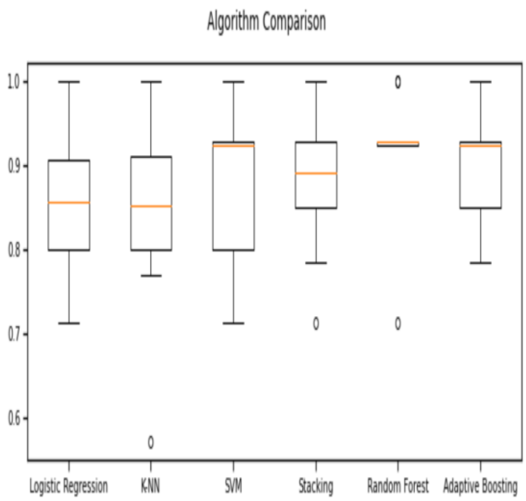

Fig. 11. Comparing model performance

Conclusion

The paper titled "PARKINSON SYNDROME PREDICTION BASED ON AUDIO MEASURES" has been done using Machine Learning techniques. The use of multiple instances learning for detecting Parkinson disease symptoms is studied. Proposed work addressed the formulation of PD symptom detection from weakly labeled data as a semi-supervised multiple instances learning problem. The features were carefully chosen to address the subject and symptom specific nature of the problem. This project shows promising preliminary results on four days of monitoring performed with two Parkinson Disease subjects. In our implementation, Parkinson disease prediction is done using machine learning algorithms. After testing the data in various algorithms like Random Forest (92%), AdaBoost (89%). The result shows that Random Forest and AdaBoost Machine achieves higher accuracy and hence both are the best fit models for predicting Parkinson Syndrome. With a larger number of patients and more complex biological data processing methodologies, this system is projected to improve in the future. This proposed Parkinson data analysis approach can also be extended to include various sorts of variances. These variations provide a wide range of results for PD detection.

References

[1] Benba, A., Jilbab, A., Hammouch, A., & Sandabad, S.(2015, March)."Voiceprints analysis using MFCC and SVM for detecting patients with Parkinson's disease".In2015 International conference on electrical and information technologies (ICEIT) (pp. 300-304).IEEE.

[2] Bhattacharya, I., & Bhatia, M. P. S. (2010, September)."SVM classification to distinguish Parkinson disease patients".In Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India (p. 14).ACM.

[3] Leandro A. Passos, "Parkinson Disease Identification using Residual Networks and Optimum-Path Forest", SACI 2018, IEEE 12th International Symposium on Applied Computational Intelligence and Informatics, May 17-19, TimiÃºoara, Romania.

[4] A. Caliskan, H. Badem, A. Basturk, and M. E. Yuksel, "Diagnosis of Parkinson disease by using deep neural network classifier," IstanbulUniversity-Journal of Electrical & Electronics Engineering, vol. 17, no. 2,pp. 3311–3318, 2017.

[5] Mathur, R., Pathak, V., &Bandil, D. (2019). "Parkinson Disease Prediction Using Machine Learning Algorithm.InEmerging Trends in Expert Applications and Security" (pp.357-363).Springer, Singapore.

[6] Rahul R. Zaveri, Prof.Pramila M.Chawan, Prediction of Parkinson's Disease using Data Mining (2020).

[7] A.Yasar et al,"Parkinson's disease prediction using Convolution neural networks"(2019).

[8] Ioannis G.Tsoulus,Georgia Mitsi"Parkinson's Disease Diagnosis Using Machine Learning and Voice".(2018).

[9] Timothy Wroge, Yasin Serdar Özkanca"Application of Machine Learning in a Parkinson's Disease Digital Biomarker Dataset Using Neural Network Construction (NNC) Methodology Discriminates Patient Motor Status",(2019).

[10] A. Lee and R. M. Gilbert, "Epidemiology of Parkinson disease," Neurologic

clinics, vol. 34, no. 4, pp. 955–965, 2016.
[11] Schalling E, Johansson K, Hartelius L. “Speech and Communication Changes Reported by People with Parkinson's Disease”, Folia PhoniatrLogop, 69(3):131-141, 2017.
[12] T. J. Wroge, Y. Özkanca, C. Demiroglu, D. Si, D. C. Atkins, and R. H.Ghomi, “Parkinson’s disease diagnosis using machine learning and voice,” IEEE Signal Processing in Medicine and Biology Symposium (SPMB). IEEE, pp. 1–7, 2018
[13] C. R. Pereira, S. A. Weber, C. Hook, G. H. Rosa, and J. P. Papa,“Deep learning-aided Parkinson’s disease diagnosis from handwritten dynamics”, 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). IEEE, pp. 340–346, 2016.
[14] ErdogduSakar, B., Isenkul, M., Sakar, C.O., Sertbas, A., Gurgen, F., Delil, S., Apaydin, H., Kursun, O, “Collection and Analysis of a Parkinson Speech Dataset with Multiple Types of Sound Recordings”, IEEE Journal of Biomedical and Health Informatics, vol. 17(4), pp. 828- 834, 2013.