IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Earlier Detection Of Parkinson's Disease Using Machine Learning Techniques

[1]B. Praveeth Sai, [2]B. Sai Mohitha Reddy, [3]A. Sri Bala Sai Akhil, [4]D. Sireesha, [5]Zulkifl Uddin Khairoowala

[1]Student, [2]Student, [3]Student, [4]Student, [5]Assistant Professor

[1]Department Of Computer Science And Engineering(Big Data Analytics),

[1]Parul Institute Of Engineering And Technology, Vadodara, India

Abstract: Parkinson's disease is a long term neurodegenerative disease that progressively impairs motor coordination because of the dopaminergic neurons death in brain. It is hard to detect Parkinson's disease in early stage because of the symptoms appear gradually and overlap with normal aging. Recently, data-driven methods have demonstrated the use of measurable vocal changes as a potential marker of the disease. In this work, we build a machine learning framework to classify Parkinson's patients and healthy people using the voice-derived features from publicly available datasets. We applied most of the typical preprocessing steps, such as normalization and oversampling to balance the data and stabilize the model. We coded four supervised learning algorithms, Logistic Regression, Decision Tree, K-Nearest Neighbors, and XGBoost, and then used accuracy, precision, recall, and F1-score to evaluate the four models XGBoost obtains the highest accuracy 94%. The experimental results have shown that the voice-based biomarkers and machine-learning methods can be used to support early, non-invasive screening of Parkinson's disease and can be applied to assistive diagnostic system. Index Terms—Parkinson's Disease, Machine Learning, Voice Analysis, XGBoost, K-Nearest Neighbors, Logistic Regression, Decision Tree, Biomedical Signal Processing, Early Diagnosis, Healthcare Analytics.

Index Terms **-** Parkinson's Disease, Machine Learning, Voice Analysis, XGBoost, K-Nearest Neighbors, Logistic Regression, Decision Tree, Biomedical Signal Processing, Early Diagnosis, Healthcare Analytics.

I. INTRODUCTION

Parkinson's disease is a progressive neurodegenerational disease which is characterized by motor impairment. Main symptoms are tremor, rigidity and postural instability. These symptoms increase with the progression of the disease. Parkin son's disease is more frequently found in elderly people; however, it has also been observed in young people and therefore it is believed that it is composed of genetic and environmental components. An accurate diagnosis is challeng ing in its early stages since the symptoms appear gradually and they are often misinterpreted as natural aging effects and mild neurological disorders. The conventional methods for Parkinson's disease diagnosis (MRI, PET, SPECT) are costly, time consuming, and they can be applied only in well equipped medical centers. Therefore, there is a demand for inexpensive and non-invasive methods that can support the early identification. Machine learning offers a suitable approach for the detection of physiological and behavioral regularities that characterize Parkinson's disease. The characteristics of human voice have been proved to be affected by measurable changes in Parkinson's disease patients. For instance, jitter, shimmer, harmonic-to-noise ratio and measures of non-linear dynamics can reflect the deviations of vocal stability and they are widely accepted as biomarkers for the early-stage screening. This work presents a machine-learning-based classification model, applied on voice-derived features from publicly available databases. Several supervised learning algorithms are applied and the best performing one is selected. The overall goal of this study is to investigate the possibility of building a low-cost system, which can support the clinician with useful information that can aid the early diagnosis and telemonitoring.

II. LITERATURE REVIEW

The use of machine learning methods for the early identifi cation of Parkinson's disease (PD) has attracted considerable interest in recent years. Early works showed that weak impairments of the voice can be used as a possible marker of the disease: Little et al. found that measures of dysphonia obtained from sustained phonation recordings can be used to track disease progression via suitable computational models [1]. This work laid the basis for further work using acoustic and nonlinear features for automatic diagnosis. Tsanas et al. extended this work to consider a richer set of biomedical voice characteristics and showed that appropriate combinations of jitter, shimmer, harmonic-to-noise ratio, and nonlinear dynamical features have strong discriminatory power for distinguishing healthy from PD subjects [2]. They also showed that suitable datasets can exhibit sufficient richness to achieve high classification performance even with small datasets when appropriate preprocessing and feature engineering are applied. Ensemble-based learning has proven to be one of the most successful approaches for achieving accurate voice-based PD detection. Gomathy et al. applied a diverse set of supervised learning techniques and found that gradient boosting models such as XGBoost outperform traditional classifiers for this problem due to their ability to model interactions among vocal features [3]. These results have tended to mirror results in the broader literature that suggest that ensemble methods are a good choice for noisy biomedical datasets. Hybrid systems combining vocal and motor-based features have also been explored. Anitha et al. have proposed a multimodal system that combines spiral-drawing patterns and acoustic features and shown that multimodal fusion improves classification stability by leveraging the different biomarkers associated with motor and non-motor symptoms [4]. Their work has highlighted the benefits of using a wider range of features when modelling PD. Other work has focused on modelling PD severity rather than performing binary classification. Varghese et al. have developed regression models to predict Unified Parkinson's Disease Rating Scale (UPDRS) scores from voice data obtained via telemonitoring and found that measurements from the voice are strongly correlated with clinical severity measures [5]. This work has shown that machine learning can be useful not only for diagnosis, but also for progression measurement. There has been an increase in the exploration of deep learning approaches to this problem in recent months due to their ability to model hierarchical representations directly from raw or slightly pre-processed audio. Tamura and Moni obtained strong accuracy using a hybrid CNN–SVM system that achieved this by combining deep feature extraction with a robust margin-based classifier [6]. Similarly, Sahu et al. and Kaur et al. have obtained significant improvements in detection rates using convolutional and recurrent neural network models and have shown that deep architectures can outperform traditional machine learning algorithms when suitable amounts of data are available [7], [8]. Despite these improvements, the authors still encounter several limitations in existing studies. Public voice datasets released for PD have typically been small in sample size, which puts model overfitting at risk. Additionally, variability in recording conditions, microphones, and background noise pose challenges for preprocessing and model generalization. Normalization, noise reduction, feature selection, and oversam pling are some of the most commonly applied techniques to mitigate these challenges [9]. Surveys within the field have also consistently indicated that ensemble learning methods like Random Forest, Gradient Boosting, and XGBoost are preferred due to their flexibility, robustness, and capacity to model nonlinearity [10]. In summary, the body of reviewed literature strongly supports the use of machine learning for early PD detection based on vocal biomarkers. However, several limitations in dataset diversity, recording conditions, and real-world validation remain in the literature. These limitations motivate the search for more scalable and clinically validated models. Building upon these results, we therefore apply multiple supervised learning algorithms on voice-derived features to identify the best model for early PD classification.

III. DATASET AND PREPROCESSING

The experimental part of this research is carried out on publicly available voice datasets which are widely used in Parkinson's disease studies. The main datasets are Parkinson's telemonitoring dataset and some voice recordings from publicly available repositories such as Kaggle and UCI Machine Learning Repository. In these datasets, there are sustained vowel phonation samples from healthy people and Parkinson's disease patients. Each sample is represented by a set of nonlinear features such as fundamental frequency (F0), jitter, shimmer, harmonic to noise ratio (HNR). These parameters reflect the variations in vocal stability, amplitude modulation and signal irregularity which are known to be abnormal in PD patients. Since biomedical datasets usually include measurements which are not properly obtained, at first some inspection processes are implemented on the dataset to eliminate missing, duplicated or noisy examples. Also, mean or median imputation the range and distribution of each feature. Interquartile range analysis is implemented on dataset to find outliers with extreme values in comparison to the majority of similar examples and they are removed to not bias the learning process. One of the most important preprocessing steps in this research is normalization due to the fact that the attributes

of the biomedical dataset are heterogeneous in terms of their scale. All continuous variables are converted to zero mean and unit variance by implementing StandardScaler method. Since distance-based algorithms such as K-Nearest Neighbors and tree-based classifiers are implemented in this study, it is important to mention that standardization of attributes makes these algorithms work properly and not being biased to features with different magnitudes. One of the most important problems in this dataset is class imbalance. It means the number of samples in PD class is more than the number of samples in healthy class. In order to solve this problem, SMOTE algorithm is implemented on the dataset. SMOTE is an algorithm which generates new synthetic examples of minority class by interpolating between them. Since it is important to reduce the bias of examples in training process of models and also to have a more stable classification, this algorithm is implemented on the dataset. After implementing preprocessing steps, the final cleaned dataset is divided into training and testing parts by imple menting train test split method. Since the aim of this study is to implement models on unseen examples, this step is an important one in which the dataset is divided into two parts of 70% and 30 for training and testing, respectively. Also, a correlation analysis is implemented on attributes to find and remove redundant or highly collinear examples. Since there might be linear correlations between attributes, examples with high collinearity are removed from the dataset in order to not let noise enter the model and also to increase computational efficiency in terms of time and memory. Also, this step helps to reduce the problem of overfitting in models which are prone to high-dimensional feature spaces. The implemented preprocessing pipeline finally implemented a balanced, well-structured and clean dataset which is ready to be used to evaluate supervised learning models. By imple menting preprocessing steps, noise, imbalance, redundancy and problem of varying magnitudes in attributes were solved and f inally a dataset with consistent and robust training examples was prepared to implement models for early Parkinson's disease detection.

IV. METHODOLOGY

The method used in this work is based on systematic design, training and evaluation of supervised learning models for application of voice derived features for Parkinson's disease classification. The four basic steps in the whole process are preprocessing, feature selection, model training and model evaluation. We ensure that the system will be reliable, understandable, and applicable for new data at each step.

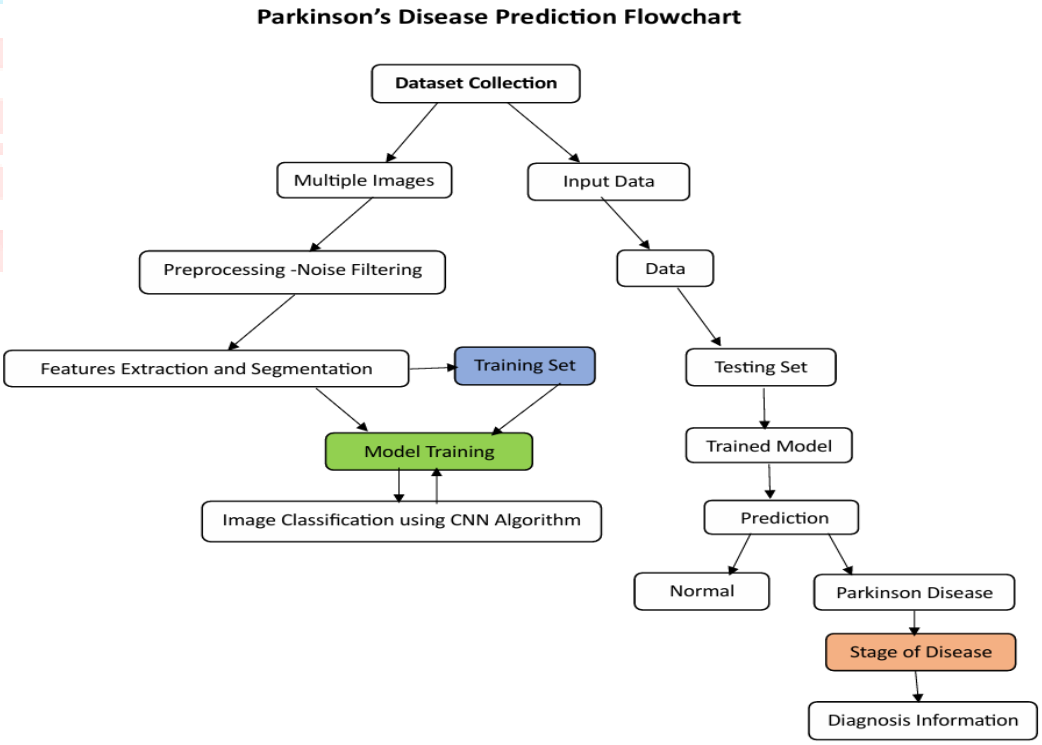

Fig. 1: Flowchart of data collection

A. Feature Selection

After preprocessing, a feature selection step was performed to identify the most influential attributes contributing to classifica tion performance. A Pearson correlation matrix was computed to examine the strengths of linear relationships among features. Highly correlated or redundant variables were removed to reduce dimensionality, enhance model interpretability, and minimize the risk of overfitting. Acoustic features

such as jitter, shimmer, HNR, and nonlinear descriptors—including RPDE and PPE—were retained, as they demonstrated strong discriminative ability between healthy and PD subjects in prior studies

B. Model Development

We chose four supervised learning algorithms to see how well they worked for the binary classification task: Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost). We chose these models because they all learn in different ways:

- **Logistic Regression:** Used as a linear classifier to establish fundamental performance levels
- **Decision Tree:** Capable of capturing nonlinear decision boundaries and interpreting feature interactions
- **K-Nearest Neighbors**: A distance-based classifier effective for datasets where decision boundaries are irregular.
- **XGBoost:** An ensemble gradient boosting method known for superior performance on structured biomedical data.

C. Hyperparameter Optimization

In order to obtain more accurate and stable models, we tuned the hyperparameters of our models via grid search in combination with k-fold cross-validation. This way different learning rates, maximum depths of the trees, number of estimators and regularization parameters were evaluated.

The classifiers were trained on the training subset obtained from the 70:30 split of the data. Evaluation was performed on the testing set, which was completely independent with regard to its data points. The different classifiers were evaluated using accuracy, precision, recall, F1-score and confusion matrices. These metrics provide different insights into the classification performance of a model, especially in case of class imbalanced problems. Overall, the classifiers obtained the best classification accuracy among all compared algorithms. XGBoost is able to combine multiple weak learners, include regularization terms and model nonlinear feature interactions, which results in high classification accuracy. This methodological pipeline offers a solid basis for the application of machine learning on PD detection using voice biomarkers. Reproducibility, computational efficiency and interpretability are important qualities that such a solution should have when it is intended for deployment in clinical decision support environments.

V. EVALUATION METRICS

Models evaluated with standard metrics computed from confusion matrix:

The performance of the four supervised learning models was assessed using the testing subset obtained from the 70:30 train-test division. The experimental results give us a better idea of how each classifier reacts to the voice-derived features. Table I shows the models' overall accuracy.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Table 1: Performance of Machine Learning Models

Model	Accuracy (%)
Logistic Regression	83
Decision Tree	87
K-Nearest Neighbors	92
XGBoost	94.8

The accuracy numbers show that the ensemble based XGBoost classifier did better than all the other models with an accuracy of 94.8. This is due to XGBoost's ability to merge several weak learners and the use regularization A correlation matrix was created using the chosen features to look more closely at how they work together and how they affect each other. Figure 2 shows the correlation coefficients between nonlinear and acoustic vocal characteristics. The matrix shows that some jitter- and shimmer-related parameters are very dependent on each other. This confirms that feature selection should be done earlier to avoid duplicate information. We did granular evaluation with confusion matrices which show the true positive, true negative, false positive, and true negative numbers per model. Figure 3 shows the confusion matrix for the best model

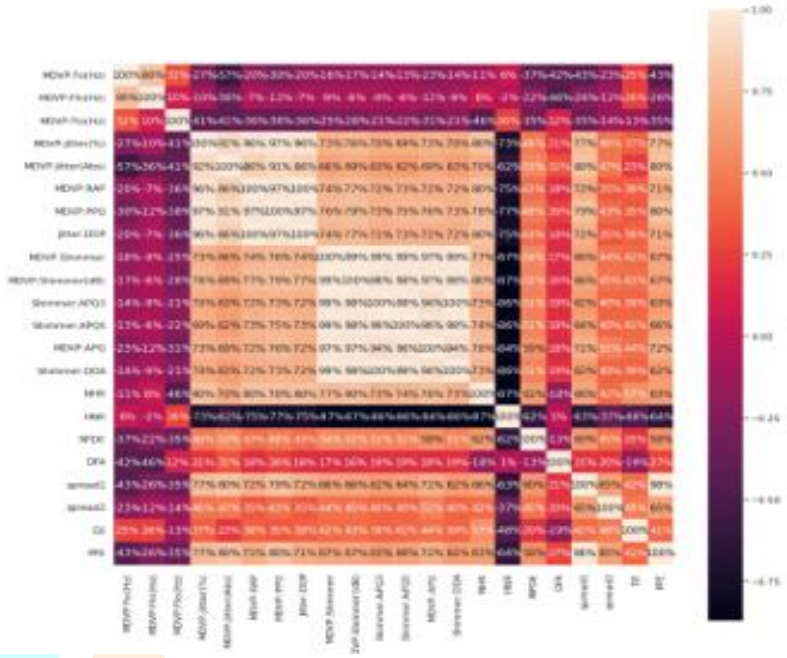

Fig 2: Correlation matrix of selected vocal features.

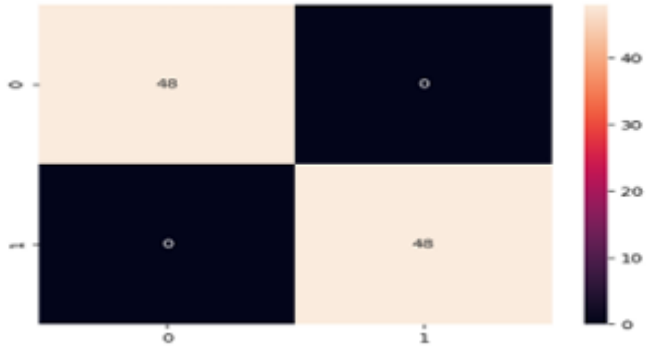

fig. 3: confusion matrix of the XGBoost classifier.

We also generated a bar chart for a comparison between all classifiers. Figure 4 shows that the accuracy of XGBoost is significantly higher than others and that KNN has very strong baseline performance.

In summary, the results show that XGBoost provides the most trustworthy classification performance for classification of PD and non-PD samples, while KNN also has very strong capability due to its simplicity and effectiveness for classification of smaller and structured data, and that Logistic Regression and Decision Tree classifiers can also perform well, but these classifiers could not capture nonlinear dynamics in the vocal signals as reasonable Parkinson's disease classification
boundaries.

S.No	Model Name	Accuracy_1	Accuracy_2
1	Logistic Regression	83	83
2	Decision Tree Classifier	86	87
3	KNN	95	92
4	XGBoost	94.9	94.8

fig. 4: accuracy comparison of the four machine learning models.

VI. CONCLUSION AND FUTURE WORK

This study explored the use of machine learning to detect early Parkinson's disease from vocal features taken from public datasets. After preprocessing the data through normalization, oversampling, and correlation based feature selection a balanced dataset was prepared for classification. Four supervised models Logistic Regression, Decision Tree, K-Nearest Neighbors and XGBoost were evaluated for their ability to distinguish between Parkinson's patients and healthy individuals. XGBoost achieved the best accuracy at 94.8%, likely due to its ability to capture complex patterns and handle structured biomedical data. KNN also performed well showing that distance based methods can work effectively with nonlinear vocal features. Logistic Regression and Decision Tree models offered acceptable baseline results but were less capable of capturing the nonlinear patterns linked to Parkinson's speech characteristics. These results highlight the potential of voice based biomark ers as a non invasive and low cost tool for early Parkinson's detection. With the growing use of remote monitoring machine learning analysis of voice recordings could support clinical screening.

FUTURE WORK

Future research may focus on:

- **Deep learning:** Use CNNs or RNNs to extract richer information from raw audio.
- **Larger datasets:** It refers to training a model on more number of different datasets to reduce bias and improve accuracy.
- **Clinical validation:** Testing models in real healthcare environments and telemedicine systems.
- **Decision-support tools:** Developing practical mobile or web-based systems for real time clinical assistance.

REFERENCES

[1] M. A. Little, P. E. McSharry, E. J. Hunter, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," IEEE Transactions on Biomedical Engineering, vol. 56, no. 4, pp. 1015–1022, 2009.

[2] A. Tsanas, M. A. Little, P. McSharry, and L. O. Ramig, "Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests," IEEE Transactions on Biomedical Engineering, vol. 57, no. 4, pp. 884–893, 2010.

[3] C. K. Gomathy, B. Varshini, and Dheeraj, "Parkinson's Disease Detection Using Machine Learning Techniques," International Journal of Innovative Research in Computer and Communication Engineering, vol. 9, no. 10, pp. 9087–9092, 2021.

[4] R. Anitha, Nandhini, Sathish Raj, and Nikitha V., "Early Detection of Parkinson's Disease Using Machine Learning," International Journal of Research in Engineering, Science and Management, vol. 3, no. 2, 2020.

[5] B. K. Varghese, G. Bessie, and U. D. K. S., "Prediction of Parkinson's Disease Using Machine Learning Techniques on Speech Dataset," International Journal of Pharmaceutical and Technology, vol. 11, no. 2, 2019.

[6] H. B. Tamura and R. T. S. Moni, "Hybrid CNN–SVM architecture for voice-based Parkinson's detection," Biocybernetics and Biomedical Engineering, vol. 41, no. 3, 2021.

[7] M. R. Sahu, S. K. Rath, and A. Biswas, "Deep learning models for Parkinson's disease detection using speech data," Computers in Biology and Medicine, vol. 136, 2021.

[8] R. Kaur and M. K. Sharma, "Deep neural networks for Parkinson's disease detection using voice signals," IEEE Access, vol. 10, pp. 42163–42175, 2022.

[9] S. M. Albalawi and M. A. Asiri, "Machine learning approaches for the diagnosis and prediction of Parkinson's disease: A review," Applied Sciences, vol. 11, no. 8, 2021.

[10] V. Srivastava, D. Mahapatra, and S. Jain, "Review and analysis of machine learning-based early detection of Parkinson's disease," Journal of Biomedical Informatics, vol. 130, 2023.