Parkinsons Disease Prediction Using ML Technique

1st Varun Gupta
Dept. of Computer Science and Engineering
KIET Group of Institutions
Ghaziabad, India
varunshivhare1729@gmail.com

2nd Yash Jain
Dept. of Computer Science and Engineering
KIET Group of Institutions
Ghaziabad, India

3rd Sumit Pal
Dept. of Computer Science and Engineering
KIET Group of Institutions
Ghaziabad, India

4th Swati Sharma
Dept. of Computer Science and Engineering
KIET Group of Institutions
Ghaziabad, India
swati.sharma@kiet.edu

Abstract: Parkinson's disease stands as the second most prevalent age-related neurological disorder, characterized by various motor and cognitive impairments. The condition presents diagnostic challenges due to its symptomatic overlap with normal aging and intentional tremors. Movement and speech difficulties typically emerge around age 50. While Parkinson's remains incurable, available medications effectively manage symptoms and enhance quality of life, making early detection crucial for slowing disease progression. This research aims to identify Parkinson's disease through speech pattern analysis using Machine Learning (ML) and Deep Learning (DL) techniques. Models were trained to distinguish between healthy individuals and those with Parkinson's disease using 195 voice samples from 31 participants sourced from the UCI machine learning repository. Performance enhancement techniques included SMOTE for data balancing, GridSearchCV for hyperparameter optimization, and feature selection methods to reduce dimensionality, improve accuracy, and prevent overfitting. Our investigation revealed superior performance from Random Forest (RF), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) algorithms. RF achieved 96.61% accuracy with 96.15% precision, while both SVM and KNN demonstrated 96.61% accuracy with perfect 100% precision scores. These findings indicate that our methodology effectively predicts Parkinson's disease and shows promise for integration into clinical diagnostic systems.

I. INTRODUCTION

Parkinson's disease (PD) represents a significant neurological disorder affecting millions globally. Its progression occurs gradually, with symptoms becoming increasingly apparent over time. While predominantly affecting individuals over 50, approximately 10% of cases manifest before age 40. The condition results from degeneration of brain nerve cells, particularly in the substantia nigra region, leading to mobility difficulties, speech impairments, and memory deterioration. Early symptoms often prove difficult to differentiate from age-related memory decline.

A 2017 economic impact study estimated PD costs in the United States at $51.9 billion, comprising $14.2 billion in direct expenses, $7.5 billion in non-medical costs, and $4.8 billion in disability-related income loss. Projections suggest these figures will reach $79 billion by 2037. Current diagnostic approaches for Parkinson's disease tend to be costly and frequently inefficient, creating demand for simpler, non-invasive detection methods.

Research indicates that voice alterations may serve as early Parkinson's indicators, positioning ML and DL models as potential diagnostic tools. This study examines how these computational approaches can identify vocal differences between healthy subjects and Parkinson's patients. We evaluated multiple models including Decision Trees, Random Forests, Logistic Regression, Support Vector Machines (SVM), Naive Bayes, K-Nearest Neighbors (KNN), and XGBoost classifiers. To determine optimal Parkinson's prediction models, we assessed performance metrics such as accuracy, F1 score, and R score. Additionally, we compared models based on interpretability, simplicity, and precision factors.

II. RELATED WORK

Previous researchers have applied machine learning techniques to Parkinson's disease detection, establishing foundations for subclassification, risk assessment, and prognosis through speech data analysis. Senturk (2020) employed Feature Importance and Recursive Feature Elimination (RFE) to identify key diagnostic attributes. Their experimentation with SVM, ANN, regression models, and classification trees demonstrated that combining SVM with RFE yielded superior accuracy at 93.84% when using vocal feature subsets.

Similarly, Gil and Manuel (2009) achieved approximately 90% accuracy in PD diagnosis using SVM and artificial neural networks. Das (2010) conducted comparative analysis of classification algorithms and neural networks, finding the latter delivered superior performance with 92.9% accuracy.

Al-Fatlawi et al. developed a deep belief network (DBN) incorporating a single output layer with two stacked Restricted Boltzmann Machines (RBMs) for Parkinson's identification. Their two-stage training approach combined supervised backpropagation learning with unsupervised RBM methods to mitigate random initial weight problems, outperforming alternative approaches with 94% accuracy.

Rasheed et al. (2020) investigated methods for enhancing Parkinson's detection through voice attribute analysis. Their fusion of Principal Component Analysis (PCA) with a backpropagation-based algorithm (BPVAM) achieved remarkable 97.50% accuracy. Another approach combining artificial neural networks with the Levenberg-Marquardt algorithm demonstrated strong performance at 95.89%.

Kadam and Jadhav (2019) developed a sparse autoencoder technique for distinguishing between Parkinson's patients and healthy controls, achieving 90% specificity with 97.28% sensitivity. A comparable deep neural network model demonstrated 90% specificity with 93.59% sensitivity.

These studies illustrate how machine learning algorithms substantially enhance speech signal analysis for Parkinson's detection, delivering improved speed and accuracy.

III. MATERIAL AND METHODS

This investigation utilized a publicly accessible dataset from the UCI Machine Learning Repository, originally collected through University of Oxford collaboration with the National Center for Voice. The dataset, designed for voice disorder research, contains recordings from 31 subjects - 8 healthy individuals (3 males, 5 females) and 23 with Parkinson's disease (16 males, 7 females). It comprises 195 total recordings with 24 distinct features extracted from voice samples. Most participants contributed six recordings, though nine subjects provided seven samples each. Participant ages ranged from 46-85 years (mean: 65.8, standard deviation: 9.8). Time since Parkinson's diagnosis varied from 0-28 years.

To maintain consistent recording quality, each 36-second audio sample was captured in a soundproof environment with microphones positioned 8 cm from the participant's mouth following standard calibration protocols. The dataset includes a "status" column where 0 indicates healthy controls and 1 signifies PD patients.

B. Strategies

Our Parkinson's disease analysis approach combines Python with Google Colab through six primary stages:

1. Information preprocessing: Initial dataset preparation includes addressing missing values, normalizing data, and organizing information for analysis.

2. Feature determination: This step enhances model effectiveness by identifying principal characteristics beneficial for Parkinson's diagnosis.

3. Engineered Minority Over-sampling Strategy (SMOTE): The generation of artificial samples for minority classes balances the dataset, reducing potential bias toward the predominant group.

4. Hyperparameter tuning (GridSearchCV): Extensive parameter testing and adjustment maximizes model accuracy and performance potential.

5. Machine and Deep Learning Models: Various classification models train on prepared data to predict Parkinson's disease presence.

6. Performance evaluation: Key performance indicators including accuracy, F1 score, precision, and recall assess model reliability and predictive capabilities.

C. Data Preprocessing

Data Preprocessing

Preprocessing represents an essential step allowing models to focus on relevant patterns while ignoring extraneous information. We loaded the dataset in CSV format into Google Colab using the Pandas library. The 'status' column showed imbalance with 147 Parkinson's disease (PD) samples and 48 healthy controls (HC) - a 75%:25% distribution.

After evaluating duplicates and missing values, we divided the dataset into 70:30 training-testing proportions to prevent underfitting or overfitting. This approach enables pattern learning from training data applicable to new information. During preprocessing, we calculated mean and standard deviation for each feature using training data to ensure proper scaling. These calculations were subsequently applied to test data through StandardScaler methodology to maintain consistent scaling.

Our research utilized several Python libraries: NumPy for scientific computing and multidimensional array operations; Pandas for data importation, reorganization, and visualization; Matplotlib and Seaborn for data visualization; and Scikit-learn (Sklearn) for machine learning processes including dimensionality reduction, clustering, regression, and classification.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

1. Feature Selection

We implemented SelectKBest to identify the dataset's top eight features. This dimensionality reduction method, used in approximately 29.1% of applications, simplifies training by selecting features with highest k scores while eliminating irrelevant data to improve usability. The eight selected features were MDVP:Fo(Hz), MDVP:Flo(Hz), MDVP:Shimmer, MDVP:APQ, HNR, spread1, spread2, and PPE.

2. Synthetic Minority Over-sampling Technique

Our dataset exhibited imbalanced distribution with more PD cases than healthy controls. While replicating existing minority class cases represents one balancing method, it provides no new information. To address this limitation, we implemented Synthetic Minority Over-sampling Technique (SMOTE), which augments minority class representation by creating synthetic cases. Rather than merely replicating actual instances, SMOTE generates new examples along line segments between minority class neighbors, promoting data diversity and enhancing model learning capability.

3. Hyperparameter tuning (GridSearchCV)

Hyperparameters constitute model parameters affecting performance that must be established before ML model training. We employed GridSearchCV for optimal hyperparameter tuning, systematically evaluating defined parameter sets, identifying best-performing combinations, and delivering expected outputs. This method proves highly effective by executing all hyperparameter combinations to produce accurate predictions.

GridSearchCV typically requires three primary inputs: 1) Estimator - the optimized machine learning model; 2) Parameter grid - hyperparameter values for estimator evaluation; and 3) Cross-Validation (CV) - dataset partitioning methodology for model assessment using K-fold validation. This technique enhances model performance by identifying optimal hyperparameters and refining predictive accuracy.

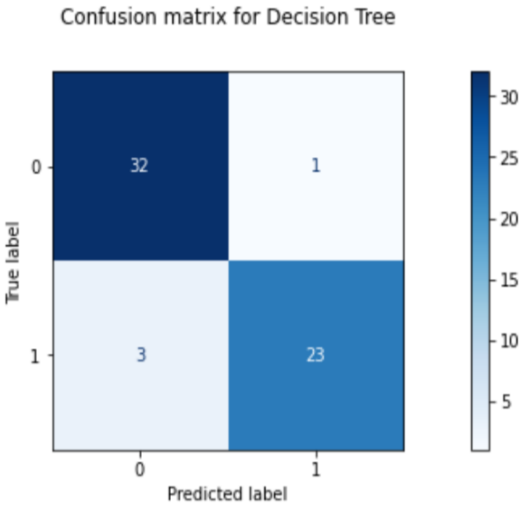
D. Classification Models

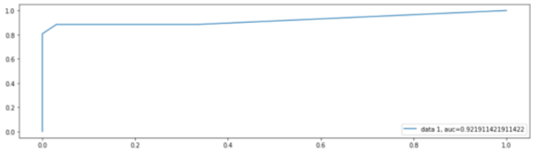
Following preprocessing, we applied numerous classifiers using ML and DL methodologies:

*1) **Decision Tree Classifier** : This effective supervised ML algorithm addresses regression and classification challenges by splitting data at each level according to different attributes. Information Gain (IG) calculations establish feature relevance for data splitting [16].*

$$IG(D,A) = H(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} \cdot H(D_v) \qquad (1)$$

where H is the entropy [16].

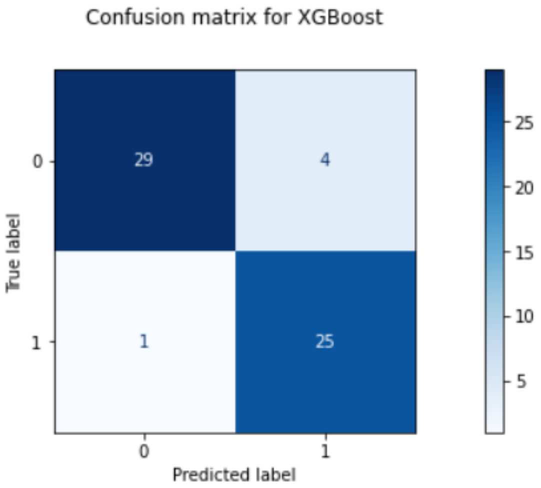




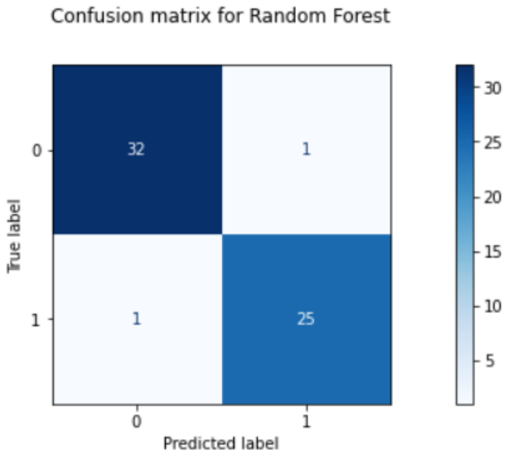
*2) **XGBoost (Extreme Gradient Boosting:** XGBoost generates multiple decision tree ensembles, with each tree addressing previous tree shortcomings. It enhances model effectiveness through gradient descent optimization, decreasing lower-level differentiable loss functions for adaptive learning.*

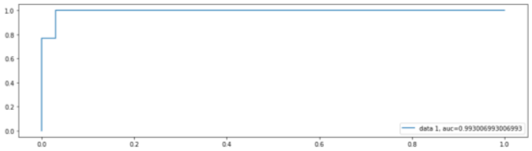
Objective Function:

$$obj(\theta) = \sum_{i}^{n} l(y_i - \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$



*3) **Random Forest Classifier:** Breiman's Random Forest methodology improves classification accuracy using multiple decision trees. Each tree utilizes randomly selected* *dataset features for classification problems. Overall decisions derive from averaging regression task outcomes or classification task voting mechanisms, introducing randomization that enhances methodology effectiveness.*

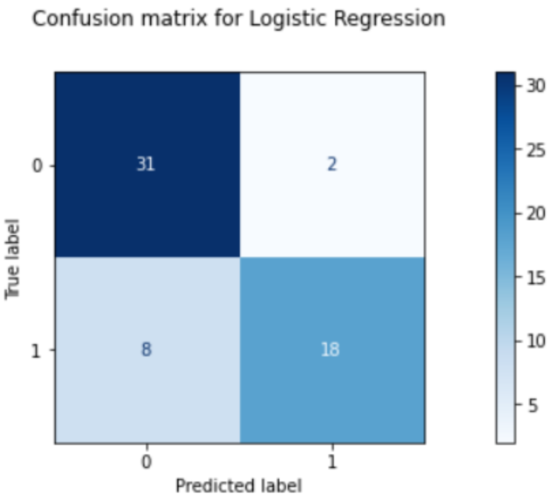


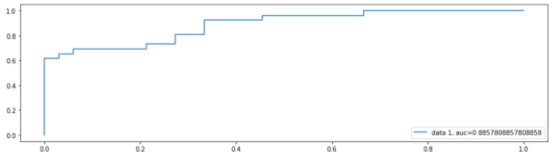


4) ***Logistic Regression***

This statistical method addresses binary classification by predicting observation categories (typically 0 or 1). Unlike linear regression's continuous outcome predictions, logistic regression determines the probability that observations belong to specific classes using the logistic/sigmoid function as a link between input features and output classes.

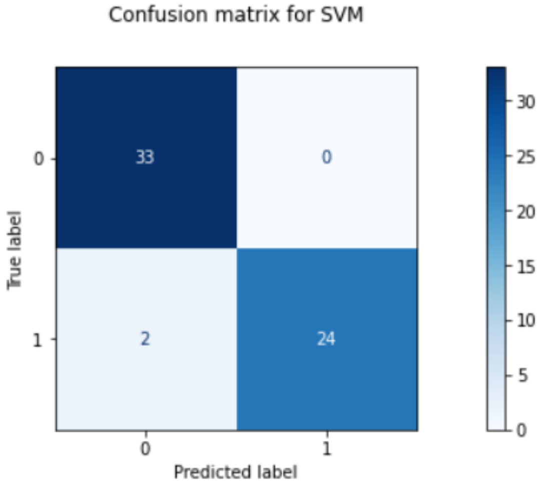
$$\ln\left(\frac{p}{1-p}\right) = y = \alpha + \beta_1 * x_1 + \beta_2 * x_2 + \cdots$$

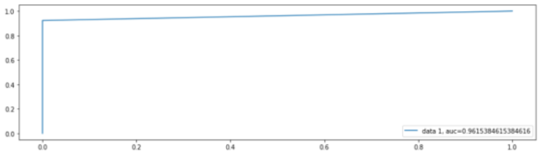




5) ***SVM***

This generic machine learning algorithm primarily serves classification purposes but accommodates regression as Support Vector Regression (SVR). SVM identifies optimal hyperplanes that accurately classify data points into individual classes, maximizing space between nearest data points and boundaries to improve predictions and generalization.

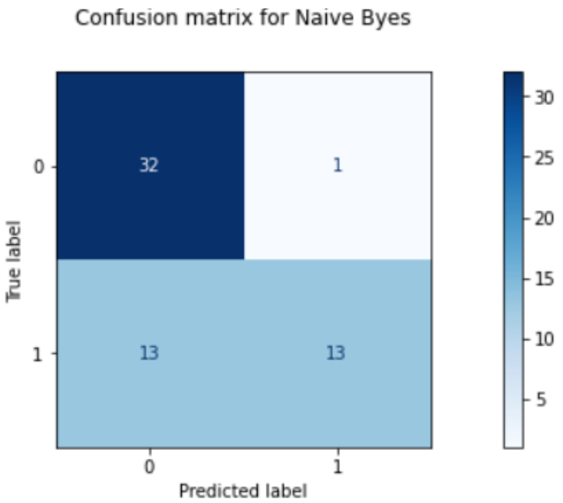


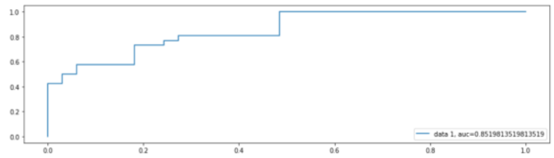


6) Naive Bayes

This classification method predicts outcomes using probability theory while assuming feature independence relative to class labels. Based on Bayes' Theorem, it performs effectively on real-world problems, particularly text classification, despite its assumption that features operate independently.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

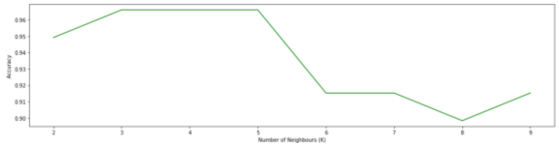


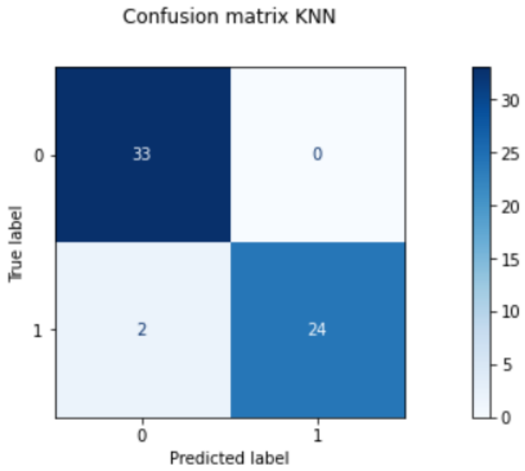


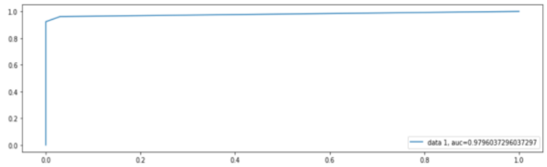
7) *KNN*

This learning algorithm examines point proximities to determine classifications. The approach requires placing similar observations together within feature space.

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$







F. *Performance Metrics*

We assessed our methodology using three key performance metrics:

1) ***Precision****: Quantifies correctly detected positive predictions relative to total positive predictions.*

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

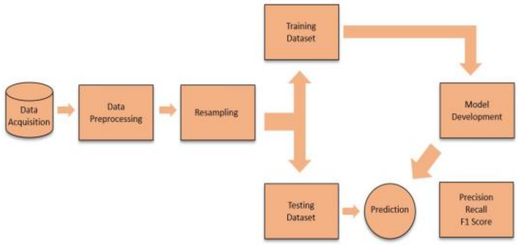


Fig. 2. Proposed Architecture

2) ***Recall:*** *Represents the ratio of correctly identified positive cases to total true positive cases.*

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

3) ***F1 Score****: Provides the harmonic mean of precision and recall, offering balanced performance assessment.*

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

These metrics effectively measure classification model performance. Precision indicates model accuracy in positive case identification, recall demonstrates completeness in finding all positive cases, and F1 score delivers balanced estimation combining both measures.

IV. PROPOSED ARCHITECTURE

In figure 2, firstly the information has been procured, at that point on this information EDA was performed and the information was preprocessed for advance work where this dataset was resampled utilizing adjusting procedures and at that point the dataset is isolated into preparing and test information and the models were prepared and the comes about were created.

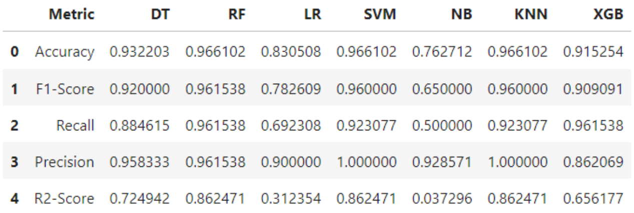
V. RESULTS AND DISCUSSIONS

This research aimed to develop an effective voice-based Parkinson's Disease diagnostic system using the UCI dataset containing 195 voice recordings from 147 PD patients and 48 healthy controls. We implemented multiple classifiers including Random Forest, Decision Tree, Support Vector Machine, k-Nearest Neighbors, and Multilayer Perceptron.

Given the dataset's imbalance with significantly more PD than HC samples, we preprocessed data using SMOTE. Additionally, we performed feature selection using SelectKBest and hyperparameter tuning through GridSearchCV. Our results demonstrate substantial performance improvements from SMOTE and GridSearchCV implementation. However, feature selection methods underperformed in this context, leading us to utilize all available features.

Recent research indicates that voice impairment often precedes motor symptoms in Parkinson's disease and can serve as an early biomarker. Our approach focuses exclusively on voice features for diagnosis, contrasting with traditional methods like UPDRS-based movement testing or DaT scans. Compared to conventional invasive approaches such as handwriting analysis or MRI scanning, our method offers faster, more precise, and cost-effective alternatives, particularly valuable since voice changes represent one of Parkinson's earliest and most apparent indicators.

Our research compared various classification models to determine the most effective methods for voice-based Parkinson's diagnosis, promoting greater accessibility and convenience.



	Metric	DT	RF	LR	SVM	NB	KNN	XGB
0	Accuracy	0.932203	0.966102	0.830508	0.966102	0.762712	0.966102	0.915254
1	F1-Score	0.920000	0.961538	0.782609	0.960000	0.650000	0.960000	0.909091
2	Recall	0.884615	0.961538	0.692308	0.923077	0.500000	0.923077	0.961538
3	Precision	0.958333	0.961538	0.900000	1.000000	0.928571	1.000000	0.862069
4	R2-Score	0.724942	0.862471	0.312354	0.862471	0.037296	0.862471	0.656177

VI. CONCLUSION & FUTURE SCOPE

Our study confirms that ML and DL methods applied to speech signal processing can accurately detect Parkinson's disease. The models outperformed conventional diagnostic approaches with remarkably high accuracy rates: Support Vector Machine (96.61%), K-nearest Neighbor (96.61%), and Random Forest (96.61%). This methodology can potentially improve patient outcomes, reduce healthcare costs, and enable earlier disease detection.

We recommend training medical practitioners and educating medical students in these techniques. With continued research and development, Parkinson's disease prediction models can become increasingly effective and beneficial worldwide.

References

[1] Al-Fatlawi A. H., Jabardi M. H., Ling S. H. (2016). “Efficient diagnosis system for parkinson's disease using deep belief network,” in *2016 IEEE Congress on evolutionary computation (CEC)* (Vancouver, BC, Canada: IEEE;), 1324–1330. 10.1109/CEC.2016.7743941 [CrossRef] [Google Scholar]

[2] Bilgen I., Guvercin G., Rekik I. (2020). Machine learning methods for brain network classification: application to autism diagnosis using cortical morphological networks. *J. Neurosci. Meth.* 343, 108799. 10.1016/j.jneumeth.2020.108799 [PubMed] [CrossRef] [Google Scholar]

[3] Bind S., Tiwari A. K., Sahani A. K., Koulibaly P., Nobili F., Pagani M., et al.. (2015). A survey of machine learning based approaches for parkinson disease prediction. *Int. J. Comput. Sci. Inf. Technol.* 6, 1648–1655. [Google Scholar]

[4] Brownlee J. (2020). *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-sensitive Learning. Machine Learning Mastery*. Available online at:

https://books.google.com.sa/books?hl=en&lr=&id=jaXJDwAAQBAJ&oi=fnd&pg=PP1&dq=Brownlee,+J.+(2020).+%E2%80%9CImbalanced+classification+with+Python:+better+metrics,+balance+skewed+classes,+cost-sensitive+learning,%E2%80%9D+in+Machine+Learning+Mastery&ots=CfNF8NM2XW&sig=6urQFaaAxqDDHzqTPTI9yjzr0rQ&redir_esc=y#v=onepage&q&f=false

[5] Charbuty B., Abdulazeez A. (2021). Classification based on decision tree algorithm for machine learning. *J. Appl. Sci. Technol. Trends.* 2, 20–28. 10.38094/jastt20165 [CrossRef] [Google Scholar]

[6] Das R. (2010). A comparison of multiple classification methods for diagnosis of parkinson disease. *Expert Syst. Appl*. 37, 1568–1572. 10.1016/j.eswa.2009.06.040 [CrossRef] [Google Scholar]

[7] Desai R. (2019). *Top 10 Python Libraries for Data Science*. Available online at: https://towardsdatascience.com/top-10-python-libraries-for-data-science-cd82294ec266 (accessed July 3, 2022).

[8] Fothergill-Misbah N., Maroo H., Hooker J., Kwasa J., Walker R. (2020). Parkinson's disease medication in kenya–situation analysis. *Pharmaceutica l J. Kenya.* 24, 38–41. [Google Scholar]

[9] Gil D., Manuel D. J. (2009). Diagnosing parkinson by using artificial neural networks and support vector machines. *Glob. J. Comput. Sci. Technol*. 9, 63–71. Available online at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.182.7856&rep=rep1&type=pdf

[10] Harel B., Cannizzaro M., Snyder P. J. (2004). Variability in fundamental frequency during speech in prodromal and incipient parkinson's disease: a longitudinal case study. *Brain Cognit.* 56, 24–29. 10.1016/j.bandc.2004.05.002 [PubMed] [CrossRef] [Google Scholar]

[11] Hossain E., Hossain M. F., Rahaman M. A. (2019). “A color and texture based approach for the detection and classification of plant leaf disease using knn classifier,” in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (Cox's Bazar: IEEE;), 1–6. 10.1109/ECACE.2019.8679247 [CrossRef] [Google Scholar]

[12] Jayaswal V. (2020). *Performance Metrics: Confusion Matrix, Precision, Recall, and f1 Score*. Available online at: https://towardsdatascience.com/performance-metrics-confusion-matrix-precision-recall-and-f1-score (accessed December 6, 2021).

[13] Kadam V. J., Jadhav S. M. (2019). “Feature ensemble learning based on sparse autoencoders for diagnosis of parkinson's disease,” in *Computing, Communication and Signal Processing. Advances in Intelligent Systems and Computing, Vol. 810*, eds B. Iyer, S. Nalbalwar, N. Pathak (Singapore: Springer;), 567–581. 10.1007/978-981-13-1513-8_58 [CrossRef] [Google Scholar]

[14] Little M. (2008). *UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science*. Available online at: https://archive.ics.uci.edu/ml/datasets/parkinsons (accessed March 17, 2023).