Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 7, No. 3, July 2025, pp: 692-712 e-ISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

A Review and Comparative Analysis on a few Classification Algorithms for Parkinson's Disease

Dhiraj Baruah[1], Rizwan Rehman[1], Pranjal Kumar Bora[1],Priyakshi Mahanta[2], Kanaka Dutta[1], Pinakshi Konwar[1]

[1] Centre for Computer Science and Applications, Dibrugarh University, Assam, India
[2] Department of Computer Applications, Jorhat Engineering College, Jorhat, India

Corresponding author: Rizwan Rehman. (e-mail: rizwan@dibru.ac.in)

Abstract **Selection and implementation of classification algorithms along with proper preprocessing methods are important for the accuracy of predictive models. This paper compares some well-known and frequently used algorithms for classification tasks and performs in depth analysis. In this study we analyzed four most frequently used algorithm viz random forest (RF), decision tree (DT), logistic regression (LR) and support vector machine (SVM). To conduct the study on the well-known Oxford Parkinson's disease Detection dataset obtained from the UCI Machine Learning Repository. We evaluated the algorithms' performance using six distinct approaches. Firstly, we used the classifiers where we didn't used any method to enhance the performance of the classifier. Secondly, we applied Principal Component Analysis (PCA) to minimize the dimensionality of the dataset. Thirdly, we used collinearity-based feature elimination (CFE) method where we applied correlation among the features and if the correlation between a pair of features exceeds the threshold of 0.9, we eliminated one from the pair. Fourthly, we adopt synthetic minority oversampling technique (SMOTE) to synthetically increase the instances of the minority class. Fifth, we combined PCA+SMOTE and on sixth method, we combined CFE + SMOTE. The study demonstrates that SVM is highly effective for Parkinson's disease classification. SVM maintained high accuracy, precision, recall and F1-score across various preprocessing techniques including PCA, CFE and SMOTE, making it robust and reliable for clinical applications. RF showed improved results with SMOTE. However, it experienced reduced performance with PCA and CFE, indicating its dependence on original feature interactions. DT benefited from PCA, while LR showed limited improvements and sensitivity to oversampling. These findings emphasize the importance of selecting appropriate preprocessing techniques to enhance model performance.**

Keywords **Classification, Decision tree, Parkinson's Disease, Random Forest, Support Vector Machine (SVM).**

I. Introduction

In classification, model predicts whether a data falls into a given category or not. It determines some appropriate mapping functions from the training dataset to predict the class label for new data entries [1]. The journey of classification algorithms begins with the paper "A Logical Calculus of the Ideas Immanent in Nervous Activity," by[2] where a mathematical model of neurons were presented, establishing the framework for artificial neural networks. This theoretical framework was the foundation for invention of the Perceptron in "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain," by [3]. It was the first technique to perform binary classification with a basic neural network model. Marvin Minsky and Seymour Papert (1969) highlighted the limits of the Perceptron, specifically its inability to tackle non-linearly separable problems[4].In the 1980s, Decision Trees emerged as a robust non-linear classification method, with[5] developing the ID3 algorithm in his work, "Induction of Decision Trees,". His work was further improved by his C4.5 algorithm [6].Cortes &Vapnik[7]introduced support vector machines (SVM) in their paper "Support-Vector Networks," showing in a substantial movement toward statistical learning in the 1990s. The kernel trick in SVM performed very well for non-linear data. Naive Bayes classifiers have also received attention, particularly in text classification, as described in earlier publications by Duda and Hart (1973). Ensemble Methods were introduced in the late 1990s and early 2000s to improve accuracy by combining numerous classifiers[8].Freund & Schapire[9] proposed AdaBoost in their work, "A

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713

Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," that created classifiers by iteratively improving weaker ones. By the same period of time [10] proposed Random Forests in his paper "Random Forests," a robust and accurate ensemble of decision trees. Finally, an upsurge of neural networks with the introduction of Deep Learning was seen, as outlined in the major study by [11]. The study demonstrated the capability of deep neural networks in challenging classification problems, paving the way for modern machine learning applications.

Classifiers in supervised learning can be broadly divided into five segments: Logical/Symbolic techniques, perceptron-based techniques, Statistical techniques, instance-based learners and Support Vector Machines.[12] described these techniques with examples. In Logical/Symbolic techniques, classifiers use decision trees, expert systems or rules to categorize data based on predetermined logic. Decision trees are example in this technique. Perceptron-Based Techniques use the perceptron model. Artificial neural network is the example of such technique. Classifiers based on statistical approaches use statistical methods to predict the correlation among input features and target labels. Common examples are Linear Discriminant Analysis (LDA), Naive Bayes and logistic regression. Classifiers in instance-based learners known as lazy learners. Their decisions depends on specific instances or examples in the training data. The most popular example is the k-NN algorithm. SVMs are the modern powerful set of supervised learning algorithms that identifies the best hyperplane for separating various classes in the feature space.

The primary aim of this study is to evaluate and compare the performance of four widely used classifiers viz., Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT) and Logistic Regression (LR)in classifying Parkinson's disease using the Oxford Parkinson's Disease dataset. The study emphasizes the role of various preprocessing techniques such as Principal Component Analysis (PCA), Collinearity based Feature Elimination (CFE) and Synthetic Minority Over-sampling Technique (SMOTE) for enhancement of classification performances. The key contributions of this work are: a comprehensive comparison of the performance of classifiers under different preprocessing configurations, an in-depth analysis of the impact of each preprocessing technique on the behavior of each classifier using precision, recall, F1-score, and ROC-AUC metrics and to identify the most reliable classifier-preprocessing combination, highlighting clinical potentials. Rest of the paper is arranged as – Section 2 presents a discussion on related work. Section 3provides an overview of the classifiers used in the experiment. Section 4 elaborates on the methodologies employed. In section 5 dataset used in the experiment is discussed. In section 6 an exploratory data analysis (EDA) performed on the dataset is discussed. Section 7 describes the classification metrics used in the experiment. Section 8 describes the experimental setup and result of the experiment. In section 9 we discussed about the limitations. Finally in section 10 we concluded our review and analysis with future work.

II. Related Work

Two feature selection algorithms- genetic algorithm (GA) and PCA were compared by [13]. SVM with GA based features gave the highest accuracy of 97.57%.[14] applying L1-norm SVM to create new subset of feature and succeeded to acquire an accuracy of 99%. [15] proposed a new multiple feature evaluation approach (MFEA) acquiring significant improvement in accuracy of many classifiers. Polar& Nour in their paper [16]classifies Parkinson's disease using a novel one against all (OGA) data sampling method. They used 45 features including pitch perturbation(4 features), Mel Frequency Cepstral Coefficients (MFCC),derivatives of aforesaid features (13 features),amplitude perturbation(5 features), harmonic-to-noise ratio(5 features), detrended fluctuation analysis (1 feature), pitch period entropy (1 feature),density entropy of recurrence period (1 features), and the ratio of glottal-to-noise excitation(1 feature).Different classifiers showed significant improvements when combined with the OGA method. When combined with OGA-II data sampling method KNN showed the highest average accuracy of 89.46 followed by SVM of average accuracy 88.76 and Logistic regression of average accuracy 84.30.[17] found that both parametric (naive bayes, logistic regression) and non-parametric (random forest, k-nearest neighbors) machine learning models can effectively classify parkinson's disease. Non-parametric models achieved higher classification accuracy compared to parametric models.[18]applied recursive feature elimination to create a better performing subset of features and acquired an accuracy of 93.84% with SVM. [19]extracted 20 audio features (13 MFCC, pitch, spectral flux, centroid, roll-off, entropy, energy, ZCR) from audio signals of four voice pathologies (laryngitis, cyst, non-fluency syndrome, and dysphonia) from the SVD dataset[20] and compare the performance of four machine learning algorithms (svm, naive bayes, decision tree, and ensemble classifier) for detection of these voice pathologies. The decision tree and naïve bayes classifiers gave the highest accuracies for detecting the voice pathologies of laryngitis, cyst, non-fluency syndrome, and dysphonia. [21] used two dimensionality reduction techniques, High Correlation Filter (HCF) and PCA to improve the performance of the classifiers and achieved 88% accuracy with SVM

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713

on “mobile parkinson disease study" dataset from the sage bionetworks mPower project[22]. Toye & Kompalli in their paper[23] classified parkinson’s disease using two datasets viz. mobile device voice recordings (MDVR- KCL) and Italian parkinson’s voice and speech database. The Italian parkinson’s voice and speech dataset has a total of 495 recordings of vowels pronunciation from 65 individuals while the MDVR-KCL dataset has 37 recordings from 37 individuals”. For both the datasets they performed three experiments. One with only acoustic features such as jitter, shimmer etc. For the second experiment they used mfcc along with the previous acoustic features and for the third experiment they utilized selected features from a combination of both acoustic and mfcc features SVM outperformed in all the three experiment of the first dataset. But for the second dataset svm showed a little poor performance compared to knn and random forest. Govindu & Palwe in their paper[24] presents a comparative analysis of several machine learning models to identify the best performing model for early detection of parkinson's disease using the Oxford university dataset. They found knn to be the best among logistic regression, svm, decision trees, extra trees classifier, k-nearest neighbors, random forests, adaboost, and gradient boosting with an accuracy of 95%. [25] used techniques such as Synthetic Minority Over-sampling Technique (SMOTE) to handle imbalanced class and hyperparameter tuning using GridSearchCV to enhance model performance. Multi-layer perceptron (MLP) and svm showed the best result to identify parkinson’s disease using features like fundamental frequency, variation in frequency, amplitude etc. [26] applied recursive feature elimination algorithm to select the features which have high positive correlation with the status. To reduce the dimention they used pca and t-SNE algorithm. Random forest achieved an accuracy of 97%. [27] combined smote-enn and svm with anrbf kernel which showed an accuracy of 96.5% in detecting parkinson's disease from speech features. The svm classifier outperforms other binary classification algorithms like random forest, k-nearest neighbours, extreme gradient boosting, decision tree, and logistic regression. [28] combined multiple existing datasets (MEEI, SVD, and a private dataset) to generate a larger "Collected and Multiple Existing Dataset (CMED)" for training and evaluating the machine learning models. They used three feature selection methods -information gain, correlation, and pca to identify the most relevant features. svm with pca method gave the highest accuracy of 99.97%. [29] gathered 12 different datasets from the parkinson's progression markers initiative (PPMI) database, which covered a range of medical evaluations, including those for motor skills, smell, cognition, sleep patterns, and depression symptoms. They identified key features, such as autonomic function, motor function, and semantic fluency as the primary contributors to Parkinson's disease.

III. Classifiers

To review the most frequently used classifying algorithms, we first searched for papers published on different supervised machine learning algorithms using Google Scholar. The data is summarized in Table 1. Since reviewing all these algorithms in a single paper would be complex and lengthy, we restrict our study to the top five algorithms from Table 1, excluding linear regression, as it is not a classifier. However, selecting classifiers for Parkinson’s disease classification requires more than just frequency; it also depends on their suitability for handling biomedical data, feature interactions, imbalanced datasets etc. Voice based Parkinson’s datasets often contain highly correlated features due to similar underlying biological signals such as vocal measurements like jitter and shimmer. svm significantly reduces the effects of high-dimensional data by determining the best hyperplane that maximizes class separation. Prior studies have reported high accuracy with SVM in detecting parkinson’s disease, often outperforming other classifiers when paired with dimensionality reduction techniques like PCA [1].Parkinson’s disease datasets often have imbalanced classes, more patients than healthy individuals. RF can handle imbalanced data better than single-tree models. RF has been successfully used in early Parkinson’s detection and has shown competitive accuracy with lower risk of overfitting compared to single decision trees[2].Since Parkinson’s disease classification involves threshold-based symptoms such as voice frequency values above or below a threshold, decision trees can model these natural decision boundaries effectively. DT is often used as a baseline model in medical diagnostics because of its explainability. Parkinson’s disease classification is fundamentally a binary task, making LR

Table 1 Comparison of Number of Papers Published for Various Supervised Algorithms

Sl. No.	Supervised ML algorithm	Number of papers published
1	Linear regression	976000
2	SVM	385000
3	Random forest	167000
4	Logistic regression	138000
5	Decision tree	125000
6	ANN	64800
7	Adaboost	33200
8	Gradient boosting machine	19500
9	Naïve bayes	18600
10	KNN	16500

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713

a natural choice. Despite its simplicity, LR has been effectively used for Parkinson's disease detection, particularly when combined with feature selection techniques like L1 regularization[4].Based on such considerations, we selected Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR), as they have been widely used in medical diagnostics and have demonstrated strong classification performance in prior research.

A. Logistic regression

Logistic regression is a binary regression which is used for classification tasks. As mentioned in[30], Let π denote the probability that Y = 1 when X = x. A logistic response function can depict the relationship between π and X. It resembles an S shaped curve. The probability π initially slowly increases with increase in X, then the increase accelerates, finally stabilises and does not exceed 1. An illustration of the S curve's shape is as follows.

The probability π initially slowly increases with increase in X, then the increase accelerates, finally stabilises and does not exceed 1. An illustration of the S curve's shape is as follows, Eq. (1)[30].

$$\pi = P(Y = 1 \mid X = x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \quad (1)$$

When we have several predictor variables,Eq. (2)[30],

$$\pi = P\left(Y = 1 \middle| X = X_1, \ldots, X_p = X_p\right) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}} \quad (2)$$

Eq. (2) [30] can be written as Eq. (3)[30],

$$1-\pi = P\left(Y = 1 \middle| X = X_1, \ldots, X_p = X_p\right) = \frac{1}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}} \quad (3)$$

Or,

Eq. (4)[30],

$$\frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p} \quad (4)$$

Taking natural logarithmic on both sides of Eq. (4) [30] we can rewrite it as Eq. (5),

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (5)$$

Here, $\frac{\pi}{1-\pi}$ is known as the odds ratio and its logarithm is known as logit. Eq (5)[30] is a linear function of the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$. The range of Eq. (5)[30] is between -∞ to +∞. In logistic regression the fitting is carried out by working with the logits[30].

B. Support vector machine

Although we can perform regression using SVM, it is mainly a classification learning model. SVM can be applied to linearly separable data adopting the mathematical equation. A kernel function can be applied on SVM to make it non-linear. The mathematical equation is. Non-linear SVM classifies by converting input feature vectors into a higher dimensional space and creating a hyperplane through them[31].

1. SVM for linearly separable data

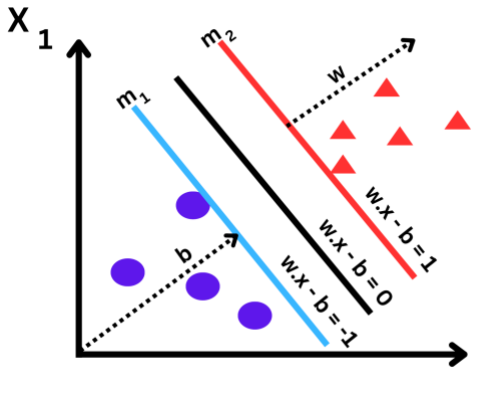

Fig. 1. Schematic diagram illustrating the separation of classes using SVM hyperplanes.

Let us consider an example of n number of set (x_1, y_1), (x_2, y_2),,(x_n,y_n). Each x_i input is associated with y_i output. Here $x_i \in R_2$ and $y_i \in$ (-1,1). Let us consider a two-dimensional space $x_i \in R_2$ as shown in Fig. 1 with a hyperplane $w^T x + b = 0$ that separates the points into two different classes. Here w represents vector perpendicular to the hyperplane and b is a scalar bias. The closest training data point to the hyperplane forms m_1 and m_1. The distance between the hyperplane and m_1, m_2 is called margin. We can draw infinite such hyperplanes but the generalization depends upon the position of the hyperplane with maximum margin [31]. The parallel hyperplanes m_1 and m_2 can be described as below in Eq. (6)[31] and Eq. (7)[32]:

$$w^T x + b_1 = 1 \quad (6)$$

$$w^T x + b_2 = -1 \quad (7)$$

From distance formula of parallel lines we can write Eq. (8)[32]

$$d = \frac{b_2 - b_1}{\sqrt{\{w^2 + 1\}}} \quad (8)$$

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713

where w is the slope of the parallel lines and $b_1 = b + 1$ and $b_2 = b - 1$ are the intercepts. Thus Eq. (9)[32],

$$d = \frac{\pm 2}{||w||} \quad (9)$$

Squaring Eq. (9)[32] we can write as in Eq. (10) [32],

$$\frac{d^2}{2} = \frac{1}{\frac{||w||^2}{2}} \quad (10)$$

To minimise the prediction error we need to maximise the distance $\frac{d^2}{2}$ or minimise $\frac{||w||^2}{2}$. This is the optimisation objective.

2. SVM with kernel

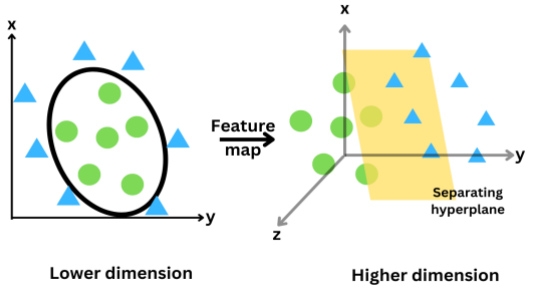

Fig. 2. SVM Feature mapping from low to high dimension in SVM with kernels.

To classify nonlinear data as shown in Fig. 2, a kernel function is used in SVM which transforms the data (x_i,y_i) of input space from a lower dimension into a higher dimensional feature space data (z_i ,y_i). A linear decision boundary in the feature space can represent a nonlinear barrier in the input space. If the transformation is appropriate, it may be possible to build a simple linear classifier in the feature space that captures the necessary nonlinearity in the input space [33]. The kernel method eliminates the need for explicit mapping when training linear learning algorithms on nonlinear functions or decision boundaries. We can write as Eq. (11)[31],

$$K : R_m \times R_m \rightarrow R \quad (11)$$

here, K is the kernel function that accepts two m-dimensional real-valued vectors and returns a real number [31]. Feature mapping is given as Eq. (12)[31],

$$k(x, x') = \langle \varphi(x). \varphi(x') \rangle_U. \quad (12)$$

here ϕ is the mapping function of x to a feature space. As explained by [33], [31]and [34]. Example of some kernel functions are:

3. Polynomial kernel

The polynomial kernel introduces non-linearity into the model by computing the similarity between vectors not just through their dot product, but by raising it to a specified degree and optionally adding a constant term. This allows the kernel to capture interactions among features up to the specified degree, making it more expressive than the linear kernel. It is particularly useful when the relationship between class labels and attributes is polynomial in nature. The flexibility and complexity of the decision boundary is significantly influenced by the choice of the degree and the constant term. A higher degree allows the model to fit more complex patterns but may also increase the risk of overfitting. The kernel is defined as Eq. (14)[31]:

$$k(x, x') = (x^t . x' + c)^d \quad (14)$$

4. Radial basis function

The RBF kernel, also known as the Gaussian kernel, is one of the most widely used kernel functions due to its ability to handle non-linear relationships effectively. Unlike linear or polynomial kernels, the RBF kernel maps input features into an infinite-dimensional space, allowing it to model very complex decision boundaries. It computes similarity based on the distance between feature vectors, where the similarity decreases with increasing distance. The RBF kernel has a key parameter gamma (γ). It controls the influence of a single training example. A small γ value implies a wider influence, while a large γ value leads to a more localized influence. Careful tuning of gamma is crucial, as inappropriate values can lead to underfitting or overfitting. The mathematical form defined as Eq. (15)[31]:

$$k(x, x') = exp\left(-\gamma \left|\left|x - x'\right|\right|^2\right) \quad (15)$$

here, γ is the kernel parameter.

C. Decision tree

In machine learning decision trees are used for classification and regression as well. The method takes the form of a treelike structure[35]. There are two kinds of nodes. The internal or decision nodes contains a condition to split the data. The leaf nodes classify a data point. The goal of a decision tree algorithm is to create a model that predicts the target variable by earning simple decision rules inferred from the data features. Decision trees are the foundational building blocks of more complex ensemble methods such as random forests and gradient boosted trees, which enhance predictive performance by combining multiple trees.Decision trees can beconstructed using the following algorithms.

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713

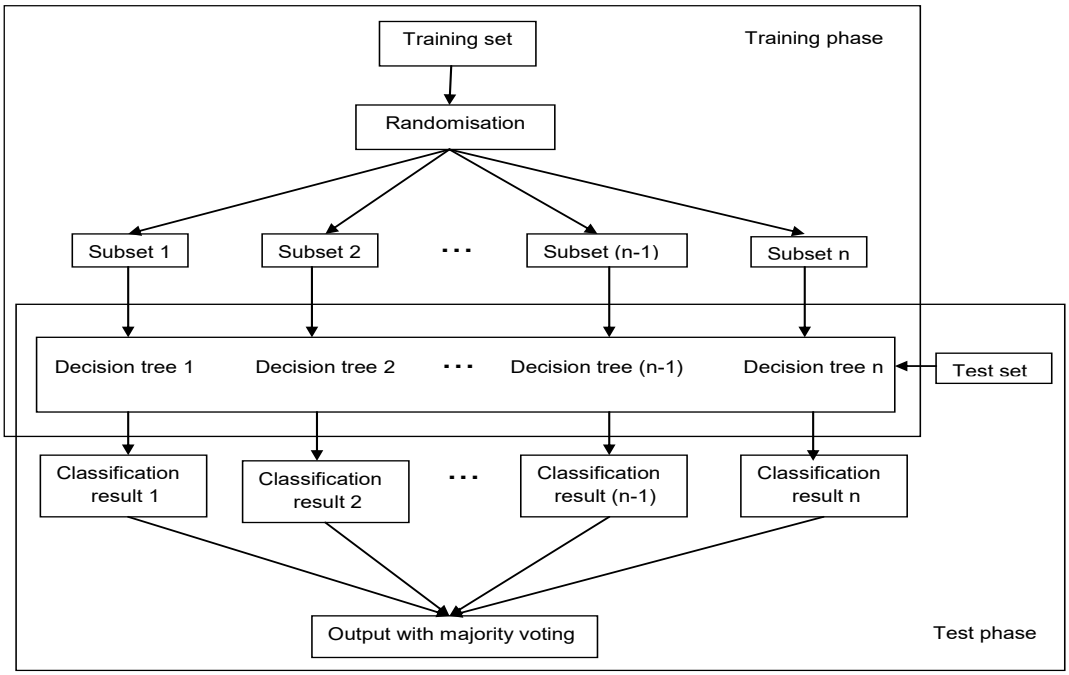

Fig. 3. Schematic diagram showing the structure of a Random Forest classifier.

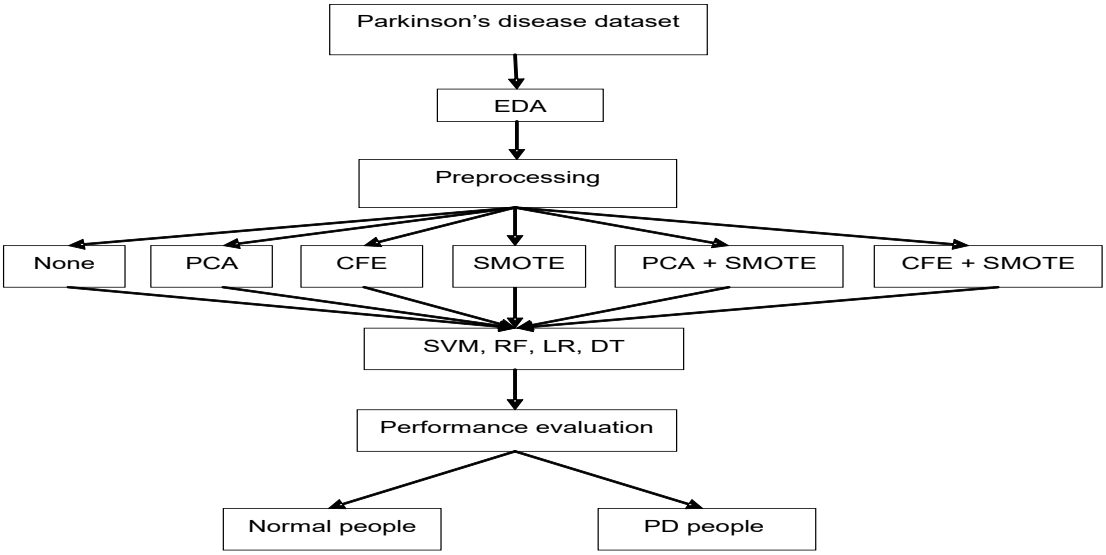

Fig. 4. Schematic diagram illustrating the steps involved in the experimental methodology.

1. id3 (Iterative DiChaudomiser 3)

The ID3 method traverses all possible tree spaces using a top-down greedy search scheme. The process begins with the complete training set and identifies the most optimal feature for the root node. The algorithm recursively calls itself with the subsets of data until it attains a leaf node with instances of the same class or there are no more features to split the data [35]. Information gain is used to get the best feature and mathematical quation can be given as Eq. (16)[35].

$$\text{Information Gain} = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \text{Entropy}(S_i) \quad (16)$$

where, Entropy(S) is the entropy of original dataset S, $|S_i|$ is the number of instances in subset $|S_i|$, $|S|$is the number of instances in the original dataset, n is the

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713

number of subsets resulting from the split.Entropy is calculated using the formula as in the Eq. (17)[35]:

$$\text{Entropy}(S) = -\sum_{i=1}^{c} p_i \log_2(p_i) \quad (17)$$

where c is the number of classes, p_i is the proportion of instance in class i in subset S.

2. CART (Classification and regression tree)

In the classification and regression tree (CART) algorithm, gini impurity is employed as the primary criterion for measuring the quality of a split, rather than entropy[35].Gini impurity quantifies the probability of incorrectly classifying a randomly selected element if it were assigned a label according to the class distribution of a given node. A Gini impurity value of zero corresponds to a perfectly pure node where all instances belong to a single class, while higher values indicate greater heterogeneity among the classes. During the construction of the decision tree, CART identifies splits that maximize the reduction in Gini impurity, thereby progressively increasing the homogeneity of the resulting subsets.

D. Random Forest

A Random Forest consists of multiple decision trees, where each decision tree is trained on a distinct subset of the training data. It selects a random subset of features for each split making its own decision based on the input data. RF introduces randomness by selecting a random subset of features at each decision split within a tree, rather than considering all available features. This feature randomness ensures that individual trees are not correlated improving the overall robustness and generalization ability of the model. Each tree in the forest independently generates a prediction based on the input data, and the final output is determined through an aggregation mechanism, typically majority voting for classification tasks or averaging for regression tasks. The schematic diagram of classification process in random forest is shown in Fig. 3[36][37].

Table 2 Feature group description highlighting the characteristics of the utilized dataset

Feature Group	Description	Example Features
Fundamental Frequency	Measures of baseline vocal frequency	MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz)
Jitter Features	Measures of frequency variation	Jitter(%), Jitter(Abs), RAP, PPQ
Shimmer Features	Measures of amplitude variation	Shimmer(dB), APQ3, APQ5
Noise-to-Harmonic Ratio (NHR) and Harmonic-to-Noise Ratio (HNR)	Measures of voice quality degradation	NHR, HNR
Nonlinear Dynamical Complexity Measures	Features capturing fractal scaling and frequency variations	RPDE, DFA, Spread1, Spread2, D2, PPE

IV. Methodology

A. Procedure

For literature review, valid repositories including Science Direct, Scopus, ACM Digital Library, IEEE Xplore, and Springer Link were used to search research papers. Besides, Google scholar search engine was also used with key words such as “supervised machine learning”, “svm”, “support vector machine”, “support vector machine with kernel”, “decision tree”, “Parkinson’s disease” etc. Filters were used on attributes like time, type etc. Papers published before 2017 were excluded. Papers published before 2024 with a few citations were excluded. For analysis of the four classifiers, we used the famous Oxford Parkinson’s disease detection dataset taken from uci machine learning repository[38] and performed exploratory data analysis (EDA) upon the dataset. After acquiring some useful information from the EDA of the dataset (which is discussed in section 4) we decided to perform the analysis in six different methods. Method 1 involved utilizing the standard classifiers directly on the original dataset without the application of any additional preprocessing or data transformation techniques, thereby serving as a baseline for performance comparison. Method 2 employed Principal Component Analysis (PCA) to address the challenges posed by high dimensionality; PCA transforms the original correlated features into a smaller set of uncorrelated principal components while preserving as much variance as possible, thus simplifying the feature space and potentially improving model efficiency and generalization. Method 3 involved the use of Collinearity-Based Feature Elimination (CFE), where features exhibiting a high correlation with other features (with a Pearson correlation coefficient greater than 0.9) were systematically removed to minimize redundancy, reduce multicollinearity, and strengthen the predictive stability of the classifiers. Method 4 applied the Synthetic Minority Oversampling Technique (SMOTE) to tackle the problem of class imbalance by synthetically generating new instances of the minority class, thereby facilitating better model learning and preventing bias toward the majority class. Method 5 combined PCA and SMOTE,

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713

wherein PCA was first applied to reduce the dimensionality and noise of the dataset, followed by SMOTE to balance the classes within the transformed feature space, ensuring comprehensive representation of both classes. Method 6 integrated CFE and SMOTE, whereby highly correlated features were initially removed through CFE, and then SMOTE was used to synthetically augment the minority class, with the objective of creating a more balanced and less redundant dataset to support more accurate and reliable classification outcomes. The schematic representation of the experimental methodology is illustrated in Fig. 4. To validate model performance, we used k-fold cross-validation. In k-fold cross-validation, the dataset is split into k folds, with each fold serving as the test set once while the rest train the model. This process repeats k times, typically using k = 5 or 10[39]. A lower value of K can lead to high variance, whereas a higher a higher value of K increases computational time complexity without significant improvement in accuracy. We used k=5.

B. Preprocessing methods

Preprocessing plays a crucial role in improving model performance by reducing dimensionality, eliminating redundancy, and addressing class imbalances. In this study, we used mainly three preprocessing techniques to enhance the models' performance viz. Principal Component Analysis (PCA), Collinearity-Based Feature Elimination (CFE), and Synthetic Minority Oversampling Technique (SMOTE). A detailed explanation of these techniques along with justification for their selection is provided below.

1. Principal Component Analysis (PCA)

PCA transforms high-dimensional data into a lower-dimensional space projecting data onto new orthogonal axes called principal components while preserving variance. These components are linear combinations of the original features[40]. The key parameter we used for PCA is number of principal components (n_components) which determines how many linear combinations of original features will be retained after transformation. It can be defined either as an integer or as a fraction representing the amount of total variance to preserve. It is common practice to keep major components that account for a considerable amount of the total variance, such as more than 80% or 90%. This approach assures that the majority of the data's information is kept while lowering dimensionality. However, excluding primary components with minimal variance can have a negative impact on classification performance[41]. we kept components that preserved at least 95% variance. Voice based Parkinson's disease datasets often have highly correlated features, leading to redundancy[42]. PCA prevents overfitting in machine learning models by lowering dimensionality while maintaining crucial information. It decreases the amount of input features, increasing computing efficiency.

2. Collinearity based Feature Elimination (CFE)

CFE removes highly correlated features to reduce redundancy and improve model interpretability[43]. Correlation threshold is the key parameter for CFE which is used to determine similar features by measuring their Pearson correlation coefficient. If the absolute value of the correlation between any two features exceeds 0.9, one of them is removed. A threshold of 0.9 effectively removes features that are nearly redundant, thereby safeguarding the retention of valuable information. Prior studies also showed that higher thresholds retain redundant features, while lower thresholds may remove too much information[44]. The choice of the correlation threshold is critical and can significantly affect model performance. Proper tuning of this parameter ensures that the model retains essential features without overfitting or underfitting.

3. Synthetic Minority Oversampling Technique (SMOTE)

SMOTE generates synthetic instances rather than duplicating existing ones to address class imbalance. This approach prevents the classifier from developing a bias towards the majority class[45]. Key parameters for SMOTE are number of nearest neighbors, sampling strategy and random state. k-nearest neighbors are used from a minority class instance to create new synthetic samples. This parameter affects the diversity of synthetic samples. Few neighbors may fail to capture true variability, whereas excessive neighbors could introduce noise from the irrelevant minority samples. Five is a frequent default that strikes a balance between having enough neighbors to represent the minority class's local structure and avoiding synthetic samples that are too similar[46].Sampling strategy is a ratio that regulates the quantity of synthetic samples produced in relation to the original minority class samples. A 1:1 ratio balances the dataset by making the minority class size equal to the majority class size. Setting a random state parameter guarantees that the same synthetic samples are produced each time the experiment is run[45]. Random oversampling, under-sampling, SMOTE with Edited Nearest Neighbors (SMOTE-ENN) are some of the alternatives to SMOTE[47]. Random oversampling duplicates existing minority class samples which may lead to overfitting by introducing duplicate instances. Under-sampling technique reduces the number of majority class instances to balance the dataset which may result in information loss. SMOTE-ENN is a hybrid preprocessing technique that combines SMOTE with data cleaning to remove noisy synthetic samples. It was not used because of concerns about losing relevant data in a small dataset.

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713

C. Expected outcomes and potential implications for clinical practices

Unlike SVM, RF and DT rely on raw feature relationships, which PCA may disrupt. PCA is a linear transformation, but tree-based models handle nonlinear relationships naturally. By transforming features, PCA may remove useful nonlinear patterns, reducing model accuracy in tree based models. When multiple features provide similar information, models may rely on unnecessary complexity rather than meaningful patterns. Thus, CFE may improve interpretability by removing highly correlated features. SMOTE increases recall by improving minority class detection but reduces precision by introducing synthetic samples that may blur class boundaries. Voice-based AI models provide a non-invasive and cost-effective screening tool enabling early detection in remote areas via telemedicine.

V. Dataset

According to National Institute of Neurological Disorders and Stroke (NINDS) Parkinson's disease (PD) is a degenerative nervous system movement disorder. Parkinson's disease occurs when brain neurons, especially those in the substantia nigra die. Dopamine is a neurotransmitter that is produced in this region and is necessary for smooth, purposeful movement. Impaired mobility results from dopamine deficiency. Typically, by the time symptoms appear, 60 to 80 percent or more of the dopamine-producing cells have already degenerated. Parkinson's patients also experience a loss of norepinephrine, which is the primary chemical transmitter to the portion of the nervous system that regulates the body's numerous automatic processes. Slowness of movement (bradykinesia), a resting tremor of the hand, stiffness (hypertonia) is among the common symptoms of Parkinson's disease. However, Diagnosis can be challenging due to the presence of various atypical symptoms, such as depression, pain, and fatigue[48]. Parkinsonian hypokinetic dysarthria, a change in speech, is one of the early signs of Parkinson's disease[49], [50]Parkinson's disease can be identified far earlier by extracting speech characteristics such as pitch, jitter, shimmer, and fundamental frequencies and feeding them into machine learning models. The dataset used in this study is the Oxford Parkinson's Disease Detection dataset, sourced from the uci machine learning repository. It comprises of several biological voice measurements from 31 individuals, 23 of whom had Parkinson's disease, was developed by [38]. A detailed analysis of the dataset used, including its size, class distribution, feature composition is mentioned below.

The dataset consists of 195 instances (samples), 22 features (excluding the target variable). Target variable (status): 1 signifies Parkinson's Disease (PD) while 0 for Healthy individuals. The Dataset was collected from 31 individuals, among whom 23 were diagnosed with Parkinson's disease, and 8 were healthy. One of the key challenges in using this dataset is its class imbalance. Approximately 75% of the samples belongs to status 1. This imbalanced distribution can lead to classifiers being biased toward the majority class. To address this, we applied the Synthetic Minority Oversampling Technique (SMOTE) to balance the class distribution. The dataset contains 22 extracted voice-based features that capture variations in fundamental frequency, amplitude, and noise-to-harmonic ratios. These features can be grouped into the following categories as shown in Table 2.

Despite being a widely used dataset for parkinson's disease classification, it has certain limitations and biases that must be considered when interpreting results. Small sample size of the dataset raises concerns about generalizability. Due to class imbalance, classifiers might show biasness towards majority class. We used SMOTE to balance the dataset before training classifiers. The dataset does not include patient demographics such as age, gender, medical history etc. Such factors could be relevant for real-world diagnosis, and their absence may limit the dataset's clinical applicability. Many features in the dataset are highly correlated. We used CFE to remove redundant features.

VI. Exploratory data analysis

We conducted an exploratory data analysis (EDA) on the dataset. The dataset has no null values. Some other interpretations of the features are described in Table 3 we examined the target variable distribution and it is found that among all the instances 75% are with status 1 (indicating person with parkinson's disease) whereas 25% are with status 0 (indicating healthy individuals). The imbalance is visually described in the Fig. 5. The correlation among the features were calculated and one of the feature-pairs which have a correlation greater than 0.9 were dropped. The feature correlation matrix is given Fig 6. After performing EDA on the dataset we can observe that the dataset is quite full of important features with no null values. But there are two issues with the dataset. First, there are many feature pairs which are highly correlated. This will create redundancy and may decrease the performance of some classifiers. Second, we also observed that the dataset faces some class imbalance. These two issues make the dataset suitable for our experiment to apply PCA, CFE and SMOTE to analyse their effect on SVM, RF, DT and LR.

VII. Classification Metrics

A. Accuracy

Accuracy is the ratio of correctly predicted observations to the total observations as in Eq. (18) [14],[17].

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad (18)$$

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713

Table 3 Presents a detailed interpretation of the features in the dataset.

Sl No.	Feature name	Description
1	MDVP: Fo (Hz)	"Average vocal fundamental frequency"
2	MDVP: Fhi (Hz)	"Maximum vocal fundamental frequency"
3	MDVP: Flo (Hz)	"Minimum vocal fundamental frequency"
4	MDVP: Jitter (%)	"Several measures of variation in fundamental frequency"
5	MDVP: Jitter (Abs)	
6	MDVP: RAP	
7	MDVP: PPQ	
8	Jitter: DDP	
9	MDVP: Shimmer	"Several measures of variation in amplitude"
10	MDVP: Shimmer (dB)	
11	Shimmer: APQ3	
12	Shimmer: APQ5	
13	MDVP: APQ	
14	Shimmer: DDA	
15	NHR	"Two measures of ratio of noise to tonal components in the voice"
16	HNR	
17	Status	"Health status of the subject (one) - Parkinson's, (zero) - healthy"
18	RPDE	"Nonlineardynamicalcomplexitymeasures"
19	DFA	"Signalfractalscalingexponent"
20	Spread1	"Nonlinear measures of fundamental frequency variation"
21	Spread2	
22	D2	"Nonlinear dynamical complexitymeasures"
23	PPE	"Nonlinear measures of fundamental frequency variation"

TP (True Positive) refers to cases where an instance with status 1 is correctly classified as PD. Conversely, TN (True Negative) represents instances where a status 0 instance is accurately identified as healthy. FP stands for false positive and is when an instance with status 0 is classified as PD. FN stands for false negative and is when an instance with status 1 is classified as healthy.

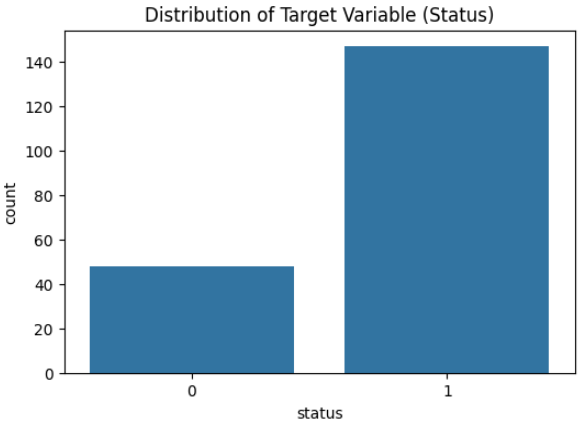

Fig. 5. Distribution of the target variable (status) showing the class balance in the dataset.

B. Precision

Precision is the proportion of correctly identified positive instances out of all instances predicted as positive as shown in the Eq. (19)[14],[17].

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (19)$$

It measures the proportion of correctly predicted PD cases out of all cases predicted as PD. High precision is very important to avoid false positive cases. False PD cases may lead to unnecessary medical tests additional medical testing, potential side effects from unwarranted treatments and anxiety for healthy individuals.

C. Recall

Recall is the proportion of correctly identified positive instances relative to the total number of actual positive instances in the class as shown in the Eq. (20)[14],[17].

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (20)$$

High recall is crucial in PD detection as missing a true PD case may lead to delayed intervention and progression of the disease without timely clinical care.

D. F1-SCORE

F1-score is the harmonic mean of precision and recall as shown in the Eq. (21)[14],[17].

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713

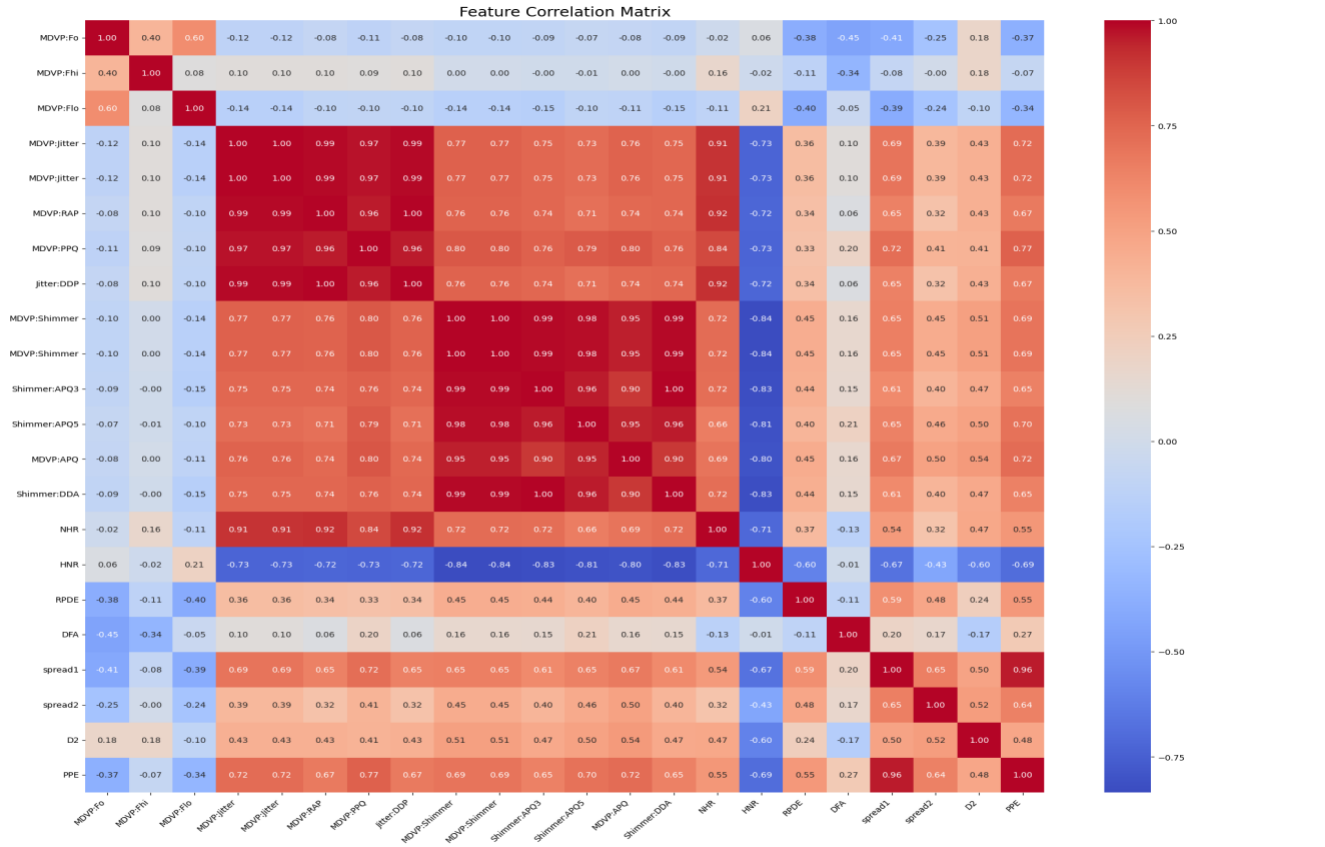

Fig. 6. Feature correlation matrix illustrating the relationships between different attributes of the dataset

$$\text{F1} - \text{score} = 2\text{x}\ \frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}} \quad (21)$$

false positives and false negatives carry risks, the F1-score is an ideal metric for evaluating performance of the classification model in Parkinson's disease.

E. Receiver Operating Characteristic - Area Under the Curve (ROC-AUC)

The Receiver Operating Characteristic (ROC) curve is a graphical representation used to evaluate the performance of a binary classification model across various threshold settings. It plots sensitivity against specificity. Sensitivity measures the proportion of actual positive cases that are correctly identified, while specificity quantifies the proportion of actual negative cases that are correctly identified. The area under the ROC curve (AUC)summarizes the overall ability of the model to discriminate between the two classes. An AUC value of 1.0 indicates perfect classification, meaning the model correctly distinguishes all positive and negative instances. An AUC of 0.5 suggests that the model has no discriminative ability and performs equivalently to random guessing.

It measures a balance between precision and recall. Achieving a high F1-score ensures that both false positives and false negatives are minimized. Since both

F. Confusion Matrix

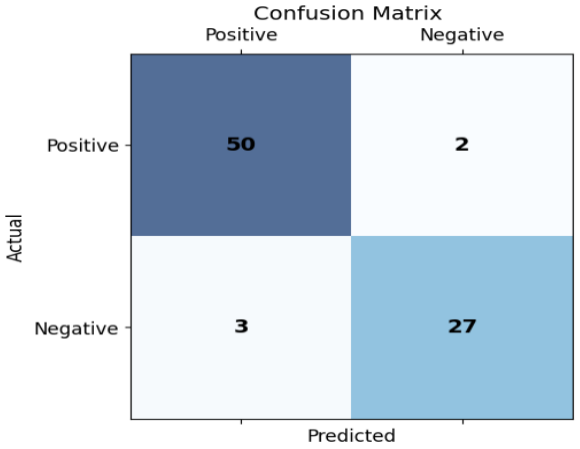

Fig. 7. Confusion matrix depicting the classification performance by showing true and false results

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713

A confusion matrix is an essential tool for evaluating classification models, particularly in medical diagnostics, where the consequences of misclassification can be severe. It compares actual vs. predicted outcomes. It summarizes the number of correct and incorrect predictions, categorized into four key components: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). For example, in Fig. 7 the model correctlypredicted50 instances as positive outcome (TP). Incorrectly predicted 2 instances as negative outcome (FN) when it was actually positive. Incorrectly predicted 3 instances a positive outcome when it was actually negative (FP) and correctly predicted 27 instances as negative outcome.

VIII. Experimental setup and results

The experiment was performed on Google colab. The runtime configurations were as – set Python 3 for runtime type and set TPU v2-8 for hardware accelerator. We used libraries like pandas, numpy, seaborn, sklearn, imlearn etc. We used k-fold cross validation with k=5 along with GridSearchCv, a tool from scikitlearn which thoroughly search for an estimator over a range of supplied parameter values. Accuracy, precision, recall and f1-score for the classifiers are shown in the Table 5. and for better interpretation graphically it is shown in Fig. 8. Fig. 9 describes ROC-AUC for each classifier and Fig. 10 describes confusion matrix of each classifier.

A. Control groups

To establish a baseline, models are trained using the original dataset that is without any enhancements (i.e. without PCA, CFE or SMOTE). This control group serves as a baseline, helping us analyze how each preprocessing method improves or degrades classifier performance. SVM, RF and LR performed well even without any preprocessing technique, suggesting that these models can handle the structure of the dataset reasonably well. DT had the lowest accuracy, likely due to feature correlation and class imbalance. Comparing preprocessed models to the control group suggests that PCA and CFE improve performance for some classifiers, while SMOTE improves results for RF but negatively impacts DT and LR.

B. Key insights from classification matrics

We can extract some of the key insights based on the experimental results with the dataset under consideration from Table 4, Fig. 8. SVM is highly suited for this classification task, especially when we reduce dimensionality either with PCA or with CFE. SVM also benefits from SMOTE. SVM maintained high performance (96.97% precision, 97.44% accuracy, 100 % recall and 98.46% f1-score) with PCA, CFE and SMOTE which suggested that SVM is robust and effective across different pre-processing techniques. RF Benefits from SMOTE but PCA and CFE reduced Its performance. It showed optimal performance with SMOTE, indicating that balancing classes may enhance RF's ensemble decision trees. Since RF uses feature interactions, it tends to perform better without dimensionality reduction, which is consistent with its relatively poorer performance when PCA is used alone. DT showed highest accuracy with PCA alone suggesting that dimensionality reduction benefits DT.LR shows less benefit from SMOTE, likely due to its sensitivity to oversampled data potentially introducing noise. Whenever SMOTE generates synthetic samples, it creates new data points that may not be as clean as the original data. SVM consistently achieved high precision reaching 96.97% in PCA, CFE, and SMOTE, indicating its strong ability to identify true PD cases without mistakenly classifying healthy cases as PD. This high precision suggests that SVM is reliable in clinical contexts where misdiagnosis of healthy individuals need to be minimized. High recall in SVM ensures that PD cases are not missed by the model. DT and LR, showed a noticeable drop in recall when combined with SMOTE, suggesting that DT and LR may be less effective at correctly identifying all PD cases under oversampled conditions. This could result from noise introduced by synthetic samples. Since PCA transforms the original features into a new set of "principal components". This makes it harder to explain how individual features affect the prediction. However, SVM with CFE or CFE+SMOTE also achieves high performance. This method balances complexity and accuracy of the model effectively, which makes it a good choice for cases where feature specific interpretation required. The ROC-AUC analysis and confusion matrix results showed that SVM with SMOTE + PCA achieved the highest AUC of 1.00, indicating reliable diagnosis with no false negatives and only one false positive making it highly suitable for medical diagnosis where missing Parkinson's cases must be minimized. RF performed well with SMOTE (AUC = 0.95), maintaining zero false negatives but slightly higher false positives, making it another strong candidate. In contrast, DT and LR struggled with recall when SMOTE was applied, making them less effective for identifying Parkinson's cases. DT performed best with CFE (AUC = 0.92), although it had more false negatives than SVM and RF which reduces its reliability. LR had the lowest AUC (0.86) and the highest number of false negatives, making it the least suitable model for this classification task.

C. Effect of each preprocessing techniques on classifier performance

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713

PCA enhanced the performance of SVM, which signifies its effectiveness in reducing feature dimensionality while preserving essential information. However, it deceased the performance of RF, likely because the model's dependency on feature interactions, which are lost after PCA transformation. CA had a positive effect on DT and a negative effect on LR, likely due to the loss of feature interpretability after transformation. Accuracy of SVM enhanced to 97.44 % from 94.87 % when we applied CFE, suggesting that removing redundant features increased its ability to classify parkinson's disease effectively. However, CFE decreased the accuracy from 94.87 % to 92.31 %, indicating feature elimination did not helped RF. Accuracy remain the same, however precision increased but recall decreased which signifies that while fewer false positives were made (higher precision), the model missed more actual positive cases (lower recall). It indicates a tradeoff between specificity and sensitivity in DT after removing correlated features. LR showed a decreased performance when redundant features were removed. SMOTE also enhanced the performance of SVM indicating that SVM benefited from a more balanced class distribution. Performance of RF remained the same with SMOTE indicating its robustness to class imbalance, likely due to its ensemble nature and ability to handle varied data distributions effectively. When SMOTE was applied on DT, precision increased a little, however accuracy and recall decreased. When SMOTE was applied on LR, it suffered a significant drop in recall, likely due to noise introduced by synthetic samples, which may have affected their decision boundaries.

Table 4 Accuracy, Precision and F1 score for some of the classifier evaluated in study.

Method	Classifiers								
	RF			LR			SVM		
	Accuracy (%)	Precision (%)	F1-score (%)	Accuracy (%)	Precision (%)	F1-score (%)	Accuracy (%)	Precision (%)	F1-score (%)
None	94.87	94.12	96.97	94.87	94.12	96.97	92.31	91.43	95.52
PCA	97.44	96.97	98.46	89.74	91.18	93.94	89.74	88.89	94.12
CFE	97.44	96.97	98.46	92.31	93.94	95.38	89.74	88.89	94.12
SMOTE	97.44	96.97	98.46	94.87	94.12	96.97	87.18	90.91	92.31
PCA + SMOTE	97.44	96.97	98.46	92.31	91.43	95.52	84.62	90.63	90.63
CFE + SMOTE	97.44	96.97	98.46	92.31	93.94	95.38	82.05	93.10	88.52

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713

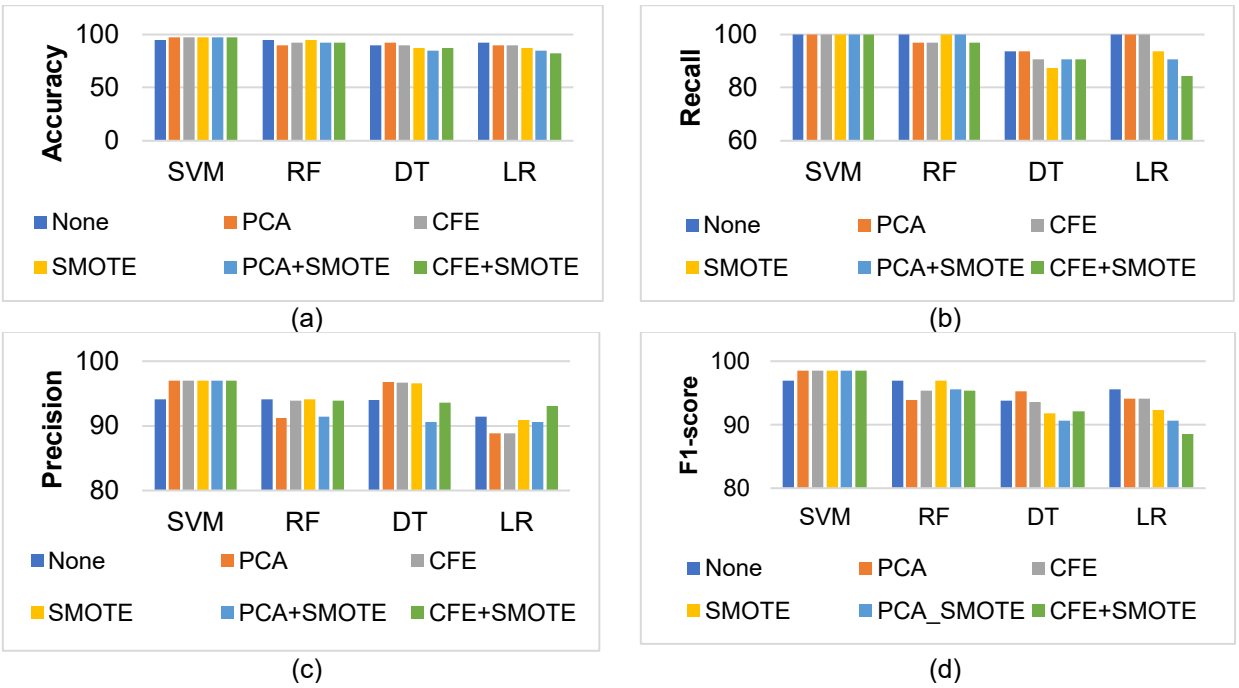

Fig. 8 Bar charts illustrating the performance of different classifiers under various preprocessing techniques: (a) Accuracy, (b) Recall, (c) Precision, and (d) F1-Score. These metrics highlight the comparative effectiveness of each classifier-preprocessing combination in classifying Parkinson's disease.

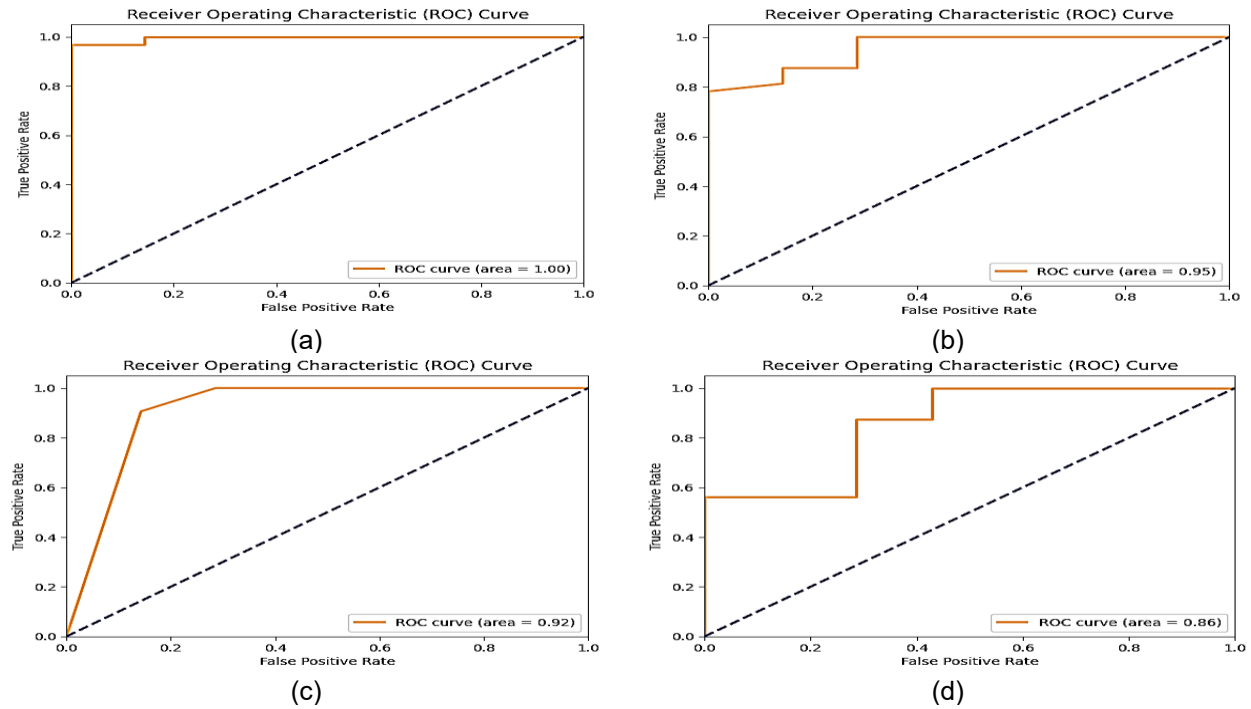

Fig. 9 ROC curves showing the classification performance of different classifier-preprocessing combinations: (a) SVM with SMOTE+PCA, (b) Random Forest with SMOTE, (c) Decision Tree with CFE, and (d) Logistic Regression with CFE+SMOTE. The curves illustrate the trade-off between true positive rate and false positive rate for each model.

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713

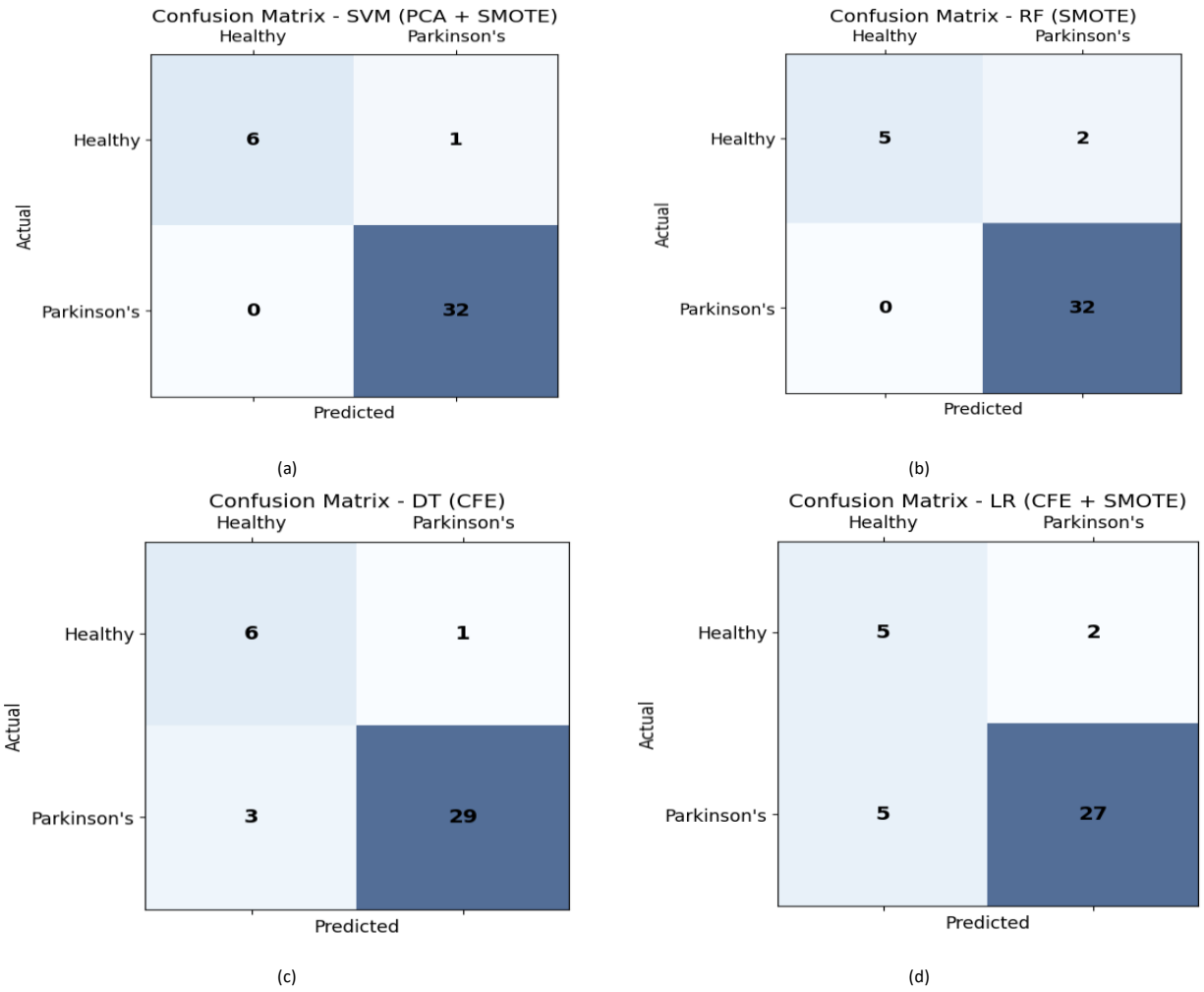

Fig. 10. Confusion matrices illustrating the classification outcomes of selected classifier-preprocessing combinations: (a) SVM with SMOTE+PCA, (b) Random Forest with SMOTE, (c) Decision Tree with CFE, and (d) Logistic Regression with CFE+SMOTE. These matrices highlight each model's performance in correctly identifying Parkinson's and healthy cases.

Some of the previous work done on the same dataset is summarized in Table 5. Aich et al. applied various classification approaches, including SVM with PCA-based and genetic algorithm-based feature selection, achieving 97.57% accuracy with SVM [13]. This aligns with our findings where SVM consistently outperforms RF, DT and LR, supporting the idea that SVM handles complex feature interactions better. Alalayah et al. used Recursive Feature Elimination (RFE) and PCA for selection of features, reporting 98% accuracy with MLP and 97% with RF [26] which complements our study, where RF and SVM showed better performance over DT and LR. KarapinarSenturk showed that SVM with Recursive Feature Elimination (RFE) achieved 93.84% accuracy [18]. Lahmiri& Shmuel optimized SVM using Bayesian optimization, achieving 92.13% accuracy with 13 selected features [51]. They used feature ranking techniques and found that removal of redundant features improved classification accuracy which supports our results that CFE improves interpretability while maintaining classifier performance. Aich et al. compared PCA and genetic algorithms for feature selection, reported that PCA was effective in reducing feature redundancy while maintaining performance. Our study builds upon this by evaluating effect of PCA on different classifiers, showing that it benefits SVM but may hinder RF and DT.

IX. Discussion

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713

This study aimed to evaluate the performance of several shallow machine learning classifiers in the context of parkinson's disease (PD) classification using voice-based features.

Our findings demonstrate that certain classifiers, particularly Support Vector Machines (SVM) and Random Forest (RF), exhibit relatively strong performance in identifying Parkinsonian patterns within the dataset [51]. This reinforces the potential of ML-based approaches for non-invasive, voice-driven diagnostic support in neurodegenerative disorders.A deeper look into the classification outcomes reveals that SVM's high accuracy likely stems from its ability to handle high-dimensional feature spaces effectively [52]. However, this benefit comes at the cost of computational expense, particularly during model training. RF, on the other hand, offered a balanced performance, robust against overfitting and more resilient in the presence of data imbalance, which is often the case in medical datasets. Its ensemble nature likely contributed to this stability. Decision Trees (DT), despite being simple and interpretable, displayed signs of overfitting, possibly due to the model's tendency to memorize training data in the absence of pruning strategies. Logistic Regression (LR), though fast and probabilistic in its output, underperformed in our setting, likely due to its linear assumptions and difficulties in managing high-dimensional and non-linearly separable voice data typical of PD patients. When compared with existing studies shown in Table 5, our results are partially consistent with previous findings. For example, the study by[53] also reported superior performance for SVM in similar PD classification tasks. However, unlike our study, they employed a larger dataset and reported even higher accuracy, which may reflect better model generalization. In contrast, our findings diverge from [54], who reported higher performance for Decision Trees, possibly due to feature engineering techniques or dataset characteristics not present in our study. Notably, while deep learning models such as CNNs and RNNs have shown promise in other studies [55], particularly when applied to raw audio or spectrogram inputs, our work chose to focus on more interpretable shallow classifiers. This decision, while deliberate, does limit direct comparability to deep learning-based studies. Despite the strengths of this work, several limitations must be acknowledged. First, the dataset used was relatively small, which inherently restricts the generalizability of our findings. The risk of overfitting increases in small datasets, particularly when models are not regularized or cross-validation is not rigorously applied. Second, the exclusive focus on shallow machine learning models, while beneficial for transparency and computational efficiency, may have prevented us from leveraging the representational power of deep learning architectures, which have recently become prominent in biomedical voice analysis. Additionally, each classifier comes with its inherent weaknesses, as discussed earlier, which influence their suitability for complex medical data such as that involved in PD classification. Beyond technical considerations, our work also touches upon broader implications. AI-driven diagnostic tools, especially those involving personal health data like voice recordings, raise important ethical and privacy concerns. Ensuring compliance with data protection standards, secure data handling practices, and informed consent are non-negotiable aspects in any real-world application. Moreover, the potential consequences of misclassification, false positives leading to undue stress or invasive procedures, and false negatives delaying essential treatment, must be carefully managed. This underlines the need for clinical validation, human-in-the-loop systems, and decision support frameworks rather than autonomous AI solutions. Overall, this research adds valuable insights to the expanding field of machine learning applications in parkinson's disease detection, emphasizing the use of interpretable and computationally efficient classifiers. Future work should consider incorporating deep learning models, expanding dataset size, and involving cross-disciplinary input from clinicians, ethicists, and data scientists to build trustworthy and effective diagnostic systems.

Table 5 Summary of some previous work done on the same dataset.

Reference	Method
[51]	The SVM classifier attained the highest classification accuracy (92.21%) using the top fourteen voice patterns selected through the Wilcoxon-based pattern ranking technique.
[18]	SVM with Recursive Feature Elimination attains accuracy of 93.84%
[26]	RF with t-distributed stochastic neighbor embedding (t-SNE) attains accuracy of 97%
[13]	Utilized two feature selection techniques, namely the genetic algorithm and principal component analysis, and applied them to various classifiers. SVM showed the highest accuracy of 97.57%

X. Conclusion

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713

This study aimed to evaluate and compare the performance of four popular classifiers viz., SVM, RF, DT and LR. to classify Parkinson's disease using the Oxford dataset. It highlighted the impact of preprocessing techniques like PCA, CFE, and SMOTE on classification performance. Key contributions of the study include thorough comparison across various preprocessing setups, detailed performance analysis using accuracy, precision, recall, F1-score and ROC-AUC and the identification of the most reliable classifier-preprocessing combination with potential clinical relevance. The experiment evaluated the effect of combining classifiers with five pre-processing methods: PCA, CFE, SMOTE, PCA + SMOTE, and CFE + SMOTE. Results indicate that SVM consistently achieved high performance (97.44% accuracy, 98.46% F1-score) across most pre-processing techniques. On the other hand, RF benefits mainly from SMOTE, showing sensitivity to dimensionality reduction methods like PCA, while LR showed less benefit from SMOTE, possibly because its sensitivity to oversampled data which may introduced noise. DT displayed performance variations based on the chosen pre-processing strategy. The study provides a broad comparison of classifiers under varied pre-processing conditions and highlights the robust performance of SVM across different pre-processing methods. Our analysis also highlights the importance of pre-processing methods, as both dimensionality reduction and oversampling significantly impact classifier accuracy and reliability. For future research we will use significance tests such as ANOVA (Analysis of Variance) to determine whether differences in classifier performance are statistically significant. We will also explore ensemble methods to mitigate individual weaknesses of classifiers, develop robust feature selection methods to reduce biases and investigate deep learning models to capture nonlinear patterns more effectively. Additionally, Parkinson's disease symptoms may vary with age, disease stage and coexisting conditions which may influence feature importance and classification accuracy. Models trained with larger and more diverse datasets will improve generalizability and reliability in real-world clinical applications.

References

[1] P. C. Sen, M. Hajra, and M. Ghosh, "Supervised Classification Algorithms in Machine Learning: A Survey and Review," in Advances in Intelligent Systems and Computing, Springer Verlag, 2020, pp. 99–111. doi: 10.1007/978-981-13-7403-6_11.

[2] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," Bull Math Biophys, vol. 5, no. 4, pp. 115–133, 1943, doi: 10.1007/BF02478259.

[3] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain.," Psychol Rev, vol. 65 6, pp. 386–408, 1958, [Online]. Available: https://api.semanticscholar.org/CorpusID:12781225

[4] M. Minsky and S. A. Papert, Perceptrons: An Introduction to Computational Geometry. The MIT Press, 2017.

[5] J. R. Quinlan, "Induction of decision trees," Mach Learn, vol. 1, no. 1, pp. 81–106, 1986, doi: 10.1007/BF00116251.

[6] S. L. Salzberg, "C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993," Mach Learn, vol. 16, no. 3, pp. 235–240, 1994, doi: 10.1007/BF00993309.

[7] C. Cortes and V. Vapnik, "Support-vector networks," Mach Learn, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[8] M. P. Sesmero, J. A. Iglesias, E. Magán, A. Ledezma, and A. Sanchis, "Impact of the learners diversity and combination method on the generation of heterogeneous classifier ensembles," Appl Soft Comput, vol. 111, p. 107689, 2021, doi: https://doi.org/10.1016/j.asoc.2021.107689.

[9] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," J Comput Syst Sci, vol. 55, no. 1, pp. 119–139, 1997, doi: https://doi.org/10.1006/jcss.1997.1504.

[10] L. Breiman, "Random Forests," Mach Learn, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[12] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," 2007.

[13] S. Aich, H.-C. Kim, K. younga, K. L. Hui, A. A. Al-Absi, and M. Sain, "A Supervised Machine Learning Approach using Different Feature Selection Techniques on Voice Datasets for Prediction of Parkinson's Disease," in 2019 21st International Conference on Advanced Communication Technology (ICACT), 2019, pp. 1116–1121.doi: 10.23919/ICACT.2019.8701961.

[14] A. U. Haqet al., "Feature Selection Based on L1-Norm Support Vector Machine and Effective Recognition System for Parkinson's Disease Using Voice Recordings," IEEE Access, vol. 7, pp. 37718–37734, 2019, doi: 10.1109/ACCESS.2019.2906350.

[15] S. A. Mostafa et al., "Examining multiple feature evaluation and classification methods for

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713

improving the diagnosis of Parkinson's disease," Cogn Syst Res, vol. 54, pp. 90–99, 2019, doi: https://doi.org/10.1016/j.cogsys.2018.12.004.

[16] K. Polat and M. Nour, "Parkinson disease classification using one against all based data sampling with the acoustic features from the speech signals," Med Hypotheses, vol. 140, p. 109678, 2020, doi: https://doi.org/10.1016/j.mehy.2020.109678.

[17] S. Sharanyaa, P. N. Renjith, and K. Ramesh, "Classification of Parkinson's Disease using Speech Attributes with Parametric and Nonparametric Machine Learning Techniques," in 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020, pp. 437–442. doi: 10.1109/ICISS49785.2020.9316078.

[18] Z. KarapinarSenturk, "Early diagnosis of Parkinson's disease using machine learning algorithms," Med Hypotheses, vol. 138, p. 109603, 2020, doi: https://doi.org/10.1016/j.mehy.2020.109603.

[19] S. A. Syed, M. Rashid, S. Hussain, A. Imtiaz, H. Abid, and H. Zahid, "Inter classifier comparison to detect voice pathologies," Mathematical Biosciences and Engineering, vol. 18, no. 3, pp. 2258–2273, 2021, doi: 10.3934/mbe.2021114.

[20] E. and O. A. and M. A. and V. J. Martínez David and Lleida, "Voice Pathology Detection on the Saarbrücken Voice Database with Calibration and Fusion of Scores Using MultiFocal Toolkit," in Advances in Speech and Language Technologies for Iberian Languages, A. and G. R. J. and H. G. L. and S. S. H. R. and R. C. D. Torre Toledano Doroteo and Ortega Giménez Alfonso and Teixeira, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 99–109.

[21] Y. C. Tai, P. G. Bryan, F. Loayza, and E. Peláez, "A voice analysis approach for recognizing Parkinson's disease patterns," IFAC-PapersOnLine, vol. 54, no. 15, pp. 382–387, 2021, doi: https://doi.org/10.1016/j.ifacol.2021.10.286.

[22] B. M. Bot et al., "The mPower study, Parkinson disease mobile data collected using ResearchKit," Sci Data, vol. 3, no. 1, p. 160011, 2016, doi: 10.1038/sdata.2016.11.

[23] A. Toye and S. Kompalli, Comparative Study of Speech Analysis Methods to Predict Parkinson's Disease. 2021. doi: 10.48550/arXiv.2111.10207.

[24] A. Govindu and S. Palwe, "Early detection of Parkinson's disease using machine learning," Procedia Comput Sci, vol. 218, pp. 249–261, 2023, doi: https://doi.org/10.1016/j.procs.2023.01.007.

[25] R. Alshammri, G. Alharbi, E. Alharbi, and I. Almubark, "Machine learning approaches to identify Parkinson's disease using voice signal features," Front ArtifIntell, vol. 6, 2023, doi: 10.3389/frai.2023.1084001.

[26] K. M. Alalayah, E. M. Senan, H. F. Atlam, I. A. Ahmed, and H. S. A. Shatnawi, "Automatic and Early Detection of Parkinson's Disease by Analyzing Acoustic Signals Using Classification Algorithms Based on Recursive Feature Elimination Method," Diagnostics, vol. 13, no. 11, 2023, doi: 10.3390/diagnostics13111924.

[27] S. Dhanalakshmi, S. Das, and R. Senthil, "Speech features-based Parkinson's disease classification using combined SMOTE-ENN and binary machine learning," Health Technol (Berl), vol. 14, no. 2, pp. 393–406, 2024, doi: 10.1007/s12553-023-00810-x.

[28] M. Ur Rehman, A. Shafique, Q.-U.-A. Azhar, S. S. Jamal, Y. Gheraibia, and A. B. Usman, "Voice disorder detection using machine learning algorithms: An application in speech and language pathology," Eng Appl ArtifIntell, vol. 133, p. 108047, 2024, doi: https://doi.org/10.1016/j.engappai.2024.108047.

[29] N. Islam, M. S. A. Turza, S. I. Fahim, and R. M. Rahman, "Single and Multi-modal Analysis for Parkinson's Disease to Detect Its Underlying Factors," Human-Centric Intelligent Systems, vol. 4, no. 2, pp. 316–334, 2024, doi: 10.1007/s44230-024-00069-z.

[30] S. Chatterjee and A. Hadi, "Regression Analysis by Example, Fourth Edition," pp. i–xvi, Apr. 2006, doi: 10.1002/0470055464.fmatter.

[31] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," Neurocomputing, vol. 408, pp. 189–215, 2020, doi: https://doi.org/10.1016/j.neucom.2019.10.118.

[32] S. Suthaharan, "Support Vector Machine," in Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, Boston, MA: Springer US, 2016, pp. 207–235. doi: 10.1007/978-1-4899-7641-3_9.

[33] R. Murty M. N. and Raghava, "Kernel-Based SVM," in Support Vector Machines and Perceptrons: Learning, Optimization, Classification, and Application to Social Networks, Cham: Springer International Publishing, 2016, pp. 57–67. doi: 10.1007/978-3-319-41063-0_5.

[34] H. P. Bhavsar and M. Panchal, "A Review on Support Vector Machine for Data Classification," 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:16365537

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713

[35] Q. and R. K. and T. X. Che Dongsheng and Liu, "Decision Tree and Ensemble Learning Algorithms with Their Applications in Bioinformatics," in Software Tools and Algorithms for Biological Systems, Q.-N. Arabnia Hamid R. and Tran, Ed., New York, NY: Springer New York, 2011, pp. 191–199. doi: 10.1007/978-1-4419-7046-6_19.

[36] L. Breiman, "Random Forests," Mach Learn, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[37] R. and P. V. Parmar Aakash and Katariya, "A Review on Random Forest: An Ensemble Classifier," in International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018, X. and L. P. and B. Z. Hemanth Jude and Fernando, Ed., Cham: Springer International Publishing, 2019, pp. 758–763.

[38] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. E. Costello, and I. M. Moroz, "Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection," Biomed Eng Online, vol. 6, no. 1, p. 23, 2007, doi: 10.1186/1475-925X-6-23.

[39] D. Krstajic, L. Buturovic, D. Leahy, and S. Thomas, "Cross-validation pitfalls when selecting and assessing regression and classification models," J Cheminform, vol. 6, p. 10, Mar. 2014, doi: 10.1186/1758-2946-6-10.

[40] I. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, p. 20150202, Apr. 2016, doi: 10.1098/rsta.2015.0202.

[41] J. Zheng and C. Rakovski, "On the Application of Principal Component Analysis to Classification Problems," Data Sci J, vol. 20, Aug. 2021, doi: 10.5334/dsj-2021-026.

[42] A. Suppaet al., "Voice in Parkinson's Disease: A Machine Learning Study," Front Neurol, vol. 13, 2022, doi: 10.3389/fneur.2022.831428.

[43] M. B. Reddy and L. S. S. Reddy, "Dimensionality Reduction: An Empirical Study on the Usability of IFE-CF (Independent Feature Elimination- by C-Correlation and F-Correlation) Measures," ArXiv, vol. abs/1002.1156, 2010, [Online]. Available: https://api.semanticscholar.org/CorpusID:10545780

[44] M. Hall, "Correlation-Based Feature Selection for Machine Learning," Department of Computer Science, vol. 19, Jun. 2000.

[45] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," J. Artif. Intell. Res. (JAIR), vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[46] T. Kosolwattana, C. Liu, R. Hu, S. Han, H. Chen, and Y. Lin, "A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare," BioData Min, vol. 16, no. 1, p. 15, 2023, doi: 10.1186/s13040-023-00330-4.

[47] H. He and E. A. Garcia, "Learning from Imbalanced Data," Knowledge and Data Engineering, IEEE Transactions on, vol. 21, pp. 1263–1284, Oct. 2009, doi: 10.1109/TKDE.2008.239.

[48] A. Bourouhou, A. Jilbab, C. Nacir, and A. Hammouch, "Comparison of classification methods to detect the Parkinson disease," in 2016 International Conference on Electrical and Information Technologies (ICEIT), 2016, pp. 421–424. doi: 10.1109/EITech.2016.7519634.

[49] N. Pah, V. Indrawati, and D. Kumar, "Voice-Based SVM Model Reliability for Identifying Parkinson's Disease," IEEE Access, vol. PP, p. 1, Sep. 2023, doi: 10.1109/ACCESS.2023.3344464.

[50] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease," J Acoust Soc Am, vol. 129, no. 1, pp. 350–367, Feb. 2011, doi: 10.1121/1.3514381.

[51] S. Lahmiri and A. Shmuel, "Detection of Parkinson's disease based on voice patterns ranking and optimized support vector machine," Biomed Signal Process Control, vol. 49, pp. 427–433, 2019, doi: https://doi.org/10.1016/j.bspc.2018.08.029.

[52] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," Neurocomputing, vol. 408, pp. 189–215, 2020, doi: https://doi.org/10.1016/j.neucom.2019.10.118.

[53] M. Wright and I. König, "Splitting on categorical predictors in random forests," PeerJ, vol. 7, p. e6339, Feb. 2019, doi: 10.7717/peerj.6339.

[54] M. Bramer, "Avoiding Overfitting of Decision Trees," in Principles of Data Mining, M. Bramer, Ed., London: Springer London, 2016, pp. 121–136. doi: 10.1007/978-1-4471-7307-6_9.

[55] R. van den Goorbergh, M. van Smeden, D. Timmerman, and B. Van Calster, "The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression," Journal of the American Medical Informatics Association, vol. 29, no. 9, pp. 1525–1534, Sep. 2022, doi: 10.1093/jamia/ocac093.

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713

Author Biography

Dr. Rizwan Rehman is an Assistant Professor at the Centre for Computer Science and Applications, Dibrugarh University, Assam, with over 20 years of academic and research experience. He holds a BCA from Jiwaji University, an MCA from RGPV Bhopal, an M.Phil from Madurai Kamaraj University, and a Ph.D. in Computer Science from Dibrugarh University. His expertise includes machine learning, data science, natural language processing (NLP), and speech processing. Dr. Rehman has contributed significantly to the development of NLP tools for low resource languages, notably the Mising and Tai languages of Northeast India. His research includes a UGC sponsored project on Mising grapheme-to-phoneme conversion and an ongoing initiative to develop NLP tools for the Tai language. He has been granted three patents by the Government of India, reflecting his innovative work in speech technology and IoT-based systems. Dr. Rehman has also developed several software applications, such as the first Mising Unicode typing tool (LÍ : SANG) and various management systems for Dibrugarh University. In addition to his research, Dr. Rehman has created a 'SWAYAM MOOCS' course on Python programming for the SWAYAM portal, which has been conducted multiple times since 2022. He actively participates in curriculum development and has mentored several Ph.D. scholars in areas like disease prediction using machine learning and speech processing. His scholarly contributions are accessible through his Scopus, ResearchGate, and Google Scholar profiles.

Dr.Pranjal Kumar Bora is an Assistant Professor at the Centre for Computer Science and Applications, Dibrugarh University, with more than 10 years of academic and research experience. He holds a Ph.D. in Computer Science from Dibrugarh University, an MCA from Gauhati University. His areas of expertise include Machine Learning, Computational Biology, Graph Theory, and Complex Networks. Mr. Bora has authored numerous research publications in high-impact journals such as Computers and Electrical Engineering. His work includes interdisciplinary collaborations and covers critical topics like Explainable AI in healthcare, amino acid network analysis, and graph-based studies on SARS-CoV-2.

Dr. Priyakshi Mahanta is an accomplished academic and researcher in the field of Computer Science, currently serving as an Assistant Professor at the Department of Computer Applications,Jorhat Engineering College, Assam. She holds a Ph.D. in Computer Science and Engineering from Tezpur University, where her research focused on bioinformatics and data mining techniques for gene expression analysis. Dr. Mahanta has more than 10 years of teaching experience at institutions including Dibrugarh University and Assam Science and Technology University, covering undergraduate and postgraduate courses. Her research interests covers bioinformatics, machine learning, data mining, and speech processing, with numerous publications in reputed journals such as Scientific Reports, Computers and Electrical Engineering, and the Journal of Biosciences. She has co-authored several conference papers and book chapters, and holds patents related to speech technology and IoT-based health devices. Dr. Mahanta is also the recipient of the Best Paper Award at the 2016 International Conference on Accessibility to Digital World and has been awarded fellowships including the UGC-BSR and NEC Merit Fellowship.

Kankana Dutta is an Assistant Professor at the Centre for Computer Science and Applications, Dibrugarh University, Assam. With a Master's degree in Computer Applications from St. Anthony's College, she has been actively engaged in academia and research since 2011. Her areas of expertise include Machine Learning and Speech Processing, with a focus on gender and language identification in low-resource languages. Before joining Dibrugarh University in 2016, she served as a faculty member at Sibsagar Commerce College, where she contributed to the departments of BCA and PGDCA. Ms. Dutta has co-authored several peer-reviewed research articles in reputed journals such as Journal of Theoretical and Applied Information Technology, Journal of Electrical Systems and Communications in Mathematics and Applications. Her recent work explores robust techniques for Tai language sentiment analysis and automatic speaker identification using neural networks and signal processing methods. In addition to her scholarly publications, she has contributed a book chapter published by Springer Nature and is a co-designer of an IoT-based alcohol breath detection device using smart sensing and deep learning. She has also delivered invited talks, including a session on email

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713

security at a digital security workshop organized by Dibrugarh University.

PinakshiKonwar is an Assistant Professor in the Centre for Computer Science and Applications, Dibrugarh University, Assam. With a strong academic foundation, she is currently pursuing her Ph.D. in Computer Science, focusing on the application of machine learning in biomedical data analysis. Her research interests span across machine learning, data mining, biomedical informatics, and artificial intelligence. She has contributed to several reputed national and international conferences. Notably, she received the Best Paper Award at the International Conference on Computational Mathematics and Applications, organized by the Department of Mathematics, NIT Silchar, for her paper titled "Potential Biomarkers for Breast Cancer Identification Using Machine Learning Techniques."

Dhiraj Baruah is a research scholar and educator specializing in Natural Language Processing (NLP), Speech Processing, and Machine Learning. Currently pursuing a Ph.D. in Computer Science from Dibrugarh University (since 2023). He hold a Master of Computer Applications from Dibrugarh University (2019) and a B.Sc. in Physics from DHSK College, Dibrugarh (2016). During his tenure as a Node.js developer intern at Assam Gas Company Ltd., Duliajan, he developed a file tracking system using barcode scanners to streamline document management, demonstrating practical proficiency in Node.js, MySQL, HTML5, and Bootstrap. His teaching experience includes serving as a faculty of Computer Science at D.H.S.K. College, Dibrugarh, instructor in COPA at Govt. ITI, Lahowal, OIL COE.

Manuscript received February 10, 2025; Revised April 27, 2025; Accepted May 10, 2025; date of publication May 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.713