Annals of Emerging Technologies in Computing (AETiC)
Vol. 10, No. 1, 2026

Research Article

Explainable AI-Assisted Parkinson's Disease Diagnosis Using Machine Learning and Deep Neural Networks

Ferdaus Ibne Aziz, Daniel Ojeda Rosales, Becky Firomssa Gudeta and Jia Uddin*

Woosong University, Republic of Korea
shifatshaheen92@gmail.com; d.ojedaro@gmail.com; becky1737@gmail.com; jia.uddin@wsu.ac.kr
*Correspondence: jia.uddin@wsu.ac.kr

Received: 23 July 2024; Accepted: 31 December 2025; Published: 1 January 2026

Abstract: Timely and specific interventions can substantially help in managing the disease, provided that the PD is diagnosed at an early stage. This paper compares machine learning (ML) and deep learning (DL) methods of PD detection with the help of vocal characteristics of a canonical sample (197 samples with 22 voice attributes pre-extracted). To reduce the issue of class imbalance, the Synthetic Minority Over-Sampling Technique (SMOTE) was used on the training data, which enhanced the strength of the models. The classical Machine Learning (ML) classifiers, such as Logistic Regression, Support Vector Machine, Random Forest, Extra Trees, Decision Tree, AdaBoost, and K-Nearest Neighbors (KNN) were evaluated, and KNN produced the best accuracy of 85% as well as competitive accuracy, recall, F1 score, and AUC ROC. In the case of deep learning, 1D CNN, 2D CNN, and LSTM were used, and 2D CNN and LSTM performed better than 1D CNN, with test accuracy of 89.7% and 84.6%, respectively, indicating their capability to learn both time-based and spatial patterns in data. Interpretability was added through Local Interpretable Model Agnostic Explanations (LIME) to ML models that point to Spread2, Recurrence Period Density Entropy, and MDVP-related frequency measures as significant vocal biomarkers. Although the framework has constraints relating to data volume and single modality, it offers a reproducible, interpretable baseline of PD detection and highlights the possibilities of explainable AI and neural networks as an assistant clinical decision-making tool. In the future, larger, multi-modal datasets are needed to offer better generalizability.

Keywords: ***Parkinson's Disease Diagnosis; Tabular Data; Machine Learning; Deep Neural Network***

1. Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder that predominantly affects the elderly population. Its prevalence continues to increase and is approaching that of Alzheimer's disease, making it one of the most common neurodegenerative conditions worldwide. PD typically manifests after the age of 60 and is characterized by motor symptoms such as rigidity, resting tremor, bradykinesia, dysphonia, and postural instability [1]. Among these manifestations, vocal impairments often emerge at an early stage of the disease, sometimes preceding more evident motor symptoms. Consequently, acoustic analysis of speech has gained increasing attention as a non-invasive and cost-effective approach for early PD detection, particularly due to the ease of collecting voice recordings in both clinical and home-based settings [2].

Despite these advances, accurate diagnosis of PD remains challenging because of substantial inter-patient variability in symptom presentation and disease progression. Clinical diagnosis primarily relies on the assessment of motor symptoms such as tremor, stiffness, and bradykinesia [3]; however, these symptoms may initially be mild or overlap with other neurological conditions. In addition, non-motor

Ferdaus Ibne Aziz, Daniel Ojeda Rosales, Becky Firomssa Gudeta and Jia Uddin, "Explainable AI-Assisted Parkinson's Disease Diagnosis Using Machine Learning and Deep Neural Networks", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 45-57, Vol. 10, No. 1, 1 January 2026, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2026.01.003, Available: https://aetic.theiaer.org/archive/v10/v10n1/p3.html.

symptoms—including sleep disturbances, cognitive impairment, and mood disorders—may appear years before motor symptoms but are frequently under-recognized in early diagnostic stages [4]. The heterogeneous progression of PD further complicates the establishment of consistent diagnostic patterns, and no single definitive clinical test currently exists. As a result, PD diagnosis depends largely on clinical judgement supported by imaging and laboratory investigations to exclude alternative disorders.

Recent advances in artificial intelligence (AI) and machine learning (ML) have demonstrated considerable potential in supporting PD identification through the analysis of speech, movement, and handwriting patterns. However, the effectiveness and generalizability of ML-based diagnostic tools are strongly influenced by dataset diversity, sample size, and data quality. The limited availability of diverse and large-scale PD datasets presents a major challenge, often restricting model robustness and clinical applicability. Prior studies [5–7] have consistently reported that a substantial proportion of individuals with PD exhibit measurable vocal abnormalities, even in early disease stages. These findings support the use of voice-based biomarkers as a practical, non-invasive, and affordable screening modality for PD. Motivated by this evidence, the present study explores ML-based analysis of vocal features for early PD detection using a well-established benchmark dataset from the UC Irvine Machine Learning Repository.

The proposed framework systematically evaluates multiple ML and DL classifiers for PD detection using vocal features. The dataset consists of 197 samples with 22 pre-extracted voice-related features. Data preprocessing includes feature scaling, correlation-based removal of redundant features, and class imbalance correction using the Synthetic Minority Over-sampling Technique (SMOTE). The dataset is partitioned into training and testing subsets using an 80:20 split. An extensive range of machine and deep learning models, such as Logistic Regression, Support Vector Machine, Random Forest, Extra Trees, Decision Tree, AdaBoost, Gradient Boosting, XGBoost, Gaussian Naive Bayes, K-Nearest Neighbor, 1D CNN, 2D CNN, and LSTM, were trained and hyperparameters optimized by grid search to maximize accuracy. Model interpretability was also adopted using Local Interpretable Model Agnostic Explanations (LIME) and permutation-based feature importance analysis to make our approach transparent.

- A comprehensive and reproducible benchmarking of classical ML and DL models for PD detection using vocal features, achieving up to 89% classification accuracy.
- Identification of clinically relevant vocal biomarkers, including Spread2, RPDE, and MDVP-related frequency measures, that contribute most significantly to PD prediction.
- Integration of explainable AI techniques to provide instance-level interpretability, supporting transparency and potential clinical adoption of ML-based diagnostic tools.

The remainder of this paper is organized as follows. Section II presents a critical review of related work on speech-based PD detection and ML methodologies. Section III details the proposed methodology, including data preprocessing, model training, and interpretability techniques. Section IV reports experimental results and comparative analyses. Finally, Section V concludes the paper with key findings, limitations, and directions for future research.

2. Related Works

The application of machine learning (ML) techniques for Parkinson's disease (PD) detection using vocal data has gained substantial attention over the past decade. Early studies primarily focused on handcrafted dysphonia features and classical classifiers to demonstrate the feasibility of speech-based PD diagnosis. Little *et al.* [8] were among the first to investigate this direction by analyzing sustained vowel phonations from a small cohort of 31 subjects. Using Support Vector Machines, they demonstrated that dysphonia measurements could discriminate between PD and healthy controls; however, the limited sample size and lack of external validation constrained the generalizability of their findings.

Subsequent research expanded on these early efforts by incorporating more advanced classifiers and optimization strategies. Ali *et al.* [9] combined L1-regularised SVM with deep neural networks and reported exceptionally high accuracies using Leave-One-Subject-Out and k-fold validation. While these results appear promising, later analyses have noted that such high performance is often influenced by strict validation protocols, limited subject diversity, and incomplete clinical metadata, which may not reflect real-world diagnostic conditions. Similarly, studies employing genetic algorithms, neural networks, and linear discriminant analysis on pre-extracted vocal features have reported accuracies exceeding 95% [13];

however, these works frequently rely on imbalanced datasets or lack transparency in feature extraction and evaluation procedures.

More recent studies have explored ensemble learning and deep learning approaches to improve robustness. Neto *et al.* [10] evaluated multiple ML and ensemble models across datasets from the UCI repository and Figshare, reporting moderate to high accuracies depending on model complexity and data source. Their findings highlight the sensitivity of model performance to dataset composition and reinforce the importance of cross-dataset validation. Other works have employed Bayesian frameworks and replicated voice recordings to address intra-subject variability [11], though limited validation and restricted dataset size remain common constraints.

Deep learning approaches have also been investigated using both canonical and large-scale clinical datasets. Wang *et al.* [14] utilized the Parkinson's Progression Markers Initiative dataset and achieved high classification accuracy using deep neural networks. Despite their strong performance, such models are often criticized for their black-box nature and limited interpretability, which pose challenges for clinical adoption. Similarly, Gunduz *et al.* [15] applied convolutional neural networks to vocal feature sets from the UCI dataset, achieving competitive accuracy but focusing primarily on single-modal inputs without extensive interpretability analysis.

Overall, the literature demonstrates that vocal biomarkers are a viable and informative modality for PD detection. However, reported performance varies widely across studies due to differences in dataset size, feature engineering strategies, validation protocols, and model transparency. In particular, many high-accuracy results are achieved under controlled experimental conditions that may not generalize to heterogeneous clinical environments. Moreover, only a limited number of studies explicitly address interpretability, despite its importance for clinician trust and decision support.

2.1. Identified Gaps

Despite notable progress in ML-based PD detection using vocal data, several key limitations persist in the existing literature. First, a significant proportion of studies rely on small, canonical, or class-imbalanced datasets, which restricts the robustness and generalizability of reported results to real-world clinical settings. Second, most approaches focus on single-modal vocal features, without considering integration with complementary modalities such as gait, tremor, or imaging data, thereby limiting diagnostic completeness.

Third, while deep learning models often achieve high predictive accuracy, they are frequently deployed as black-box systems with minimal interpretability, making them difficult to translate into clinical practice. Finally, comparative evaluations across multiple ML models using consistent preprocessing, validation, and evaluation metrics are often lacking, complicating fair benchmarking and reproducibility

Motivated by these gaps, the present study aims to provide a transparent and reproducible benchmarking framework for PD detection using vocal features. By systematically evaluating multiple classical ML models, explicitly addressing class imbalance, and integrating explainable AI techniques for instance-level interpretation, this work seeks to bridge the gap between predictive performance and clinical interpretability while acknowledging the limitations imposed by dataset size and modality.

3. Methodology

Figure 1 shows that the proposed workflow in the study of Parkinson's disease diagnosis is based on data preprocessing, feature selection, machine learning, and deep learning model training, and interpretability analysis to deliver a precise and transparent outcome.

3.1. Dataset Description

The dataset, sourced from the UC Irvine ML Repository [8], is multivariate, encompassing 197 instances across 22 features. It comprises data from 31 individuals, 23 of whom are diagnosed with PD, and includes diverse biological voice measurements. Each column in the dataset represents a distinct voice metric, while each row corresponds to one of the 195 voice recordings from these individuals, identified in the "name" column. The "status" column is binary-coded, with 0 indicating healthy subjects and 1 indicating those with PD. The dataset's primary objective is to facilitate the differentiation between healthy individuals and those afflicted with the disease. Table 1 presents the feature names and their descriptions.

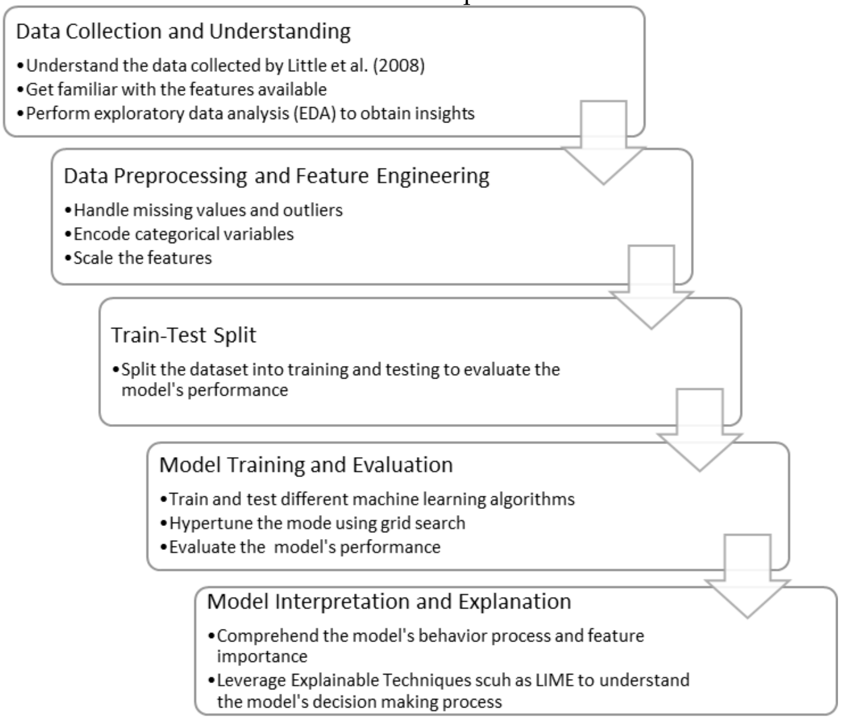

Data Collection and Understanding
•Understand the data collected by Little et al. (2008)
•Get familiar with the features available
•Perform exploratory data analysis (EDA) to obtain insights
Data Preprocessing and Feature Engineering
•Handle missing values and outliers
•Encode categorical variables
•Scale the features
Train-Test Split
•Split the dataset into training and testing to evaluate the model's performance
Model Training and Evaluation
•Train and test different machine learning algorithms
•Hypertune the mode using grid search
•Evaluate the model's performance
Model Interpretation and Explanation
•Comprehend the model's behavior process and feature importance
•Leverage Explainable Techniques scuh as LIME to understand the model's decision making process

Figure 1. Proposed Methodology

Table 1. Features names and their descriptions.

Feature Name	Description
MDVP: Fo (Hz)	Average vocal fundamental frequency
MDVP: Fhi (Hz)	Maximum vocal fundamental frequency
MDVP: Flo (Hz)	Minimum vocal fundamental frequency
MDVP: Jitter (%)	variation in the duration of the voice cycles, measured as a percentage
MDVP: Jitter (Abs)	variation in the duration of the voice cycles, measured in absolute values
MDVP: RAP	Relative Amplitude Perturbation, a measure of the variation in the amplitude of the voice
MDVP: PPQ	Pitch Period Perturbation Quotient, a measure of the variation in the amplitude of the voice
Jitter: DDP	Average Absolute Difference of Differences, a measure of the variation in the duration of the voice cycles
MDVP: Shimmer	Shimmer (local). The variation in the amplitude of the voice, related to the roughness of the voice
MDVP: Shimmer(dB)	Shimmer (decibels). The variation in the amplitude of the voice, measured in decibels
Shimmer: APQ3	Shimmer (three-point amplitude perturbation quotient), a measure of the variation in the amplitude of the voice
Shimmer: APQ5	Shimmer (five-point amplitude perturbation quotient), a measure of the variation in the amplitude of the voice
MDVP: APQ	Shimmer (amplitude perturbation quotient), a measure of the variation in the amplitude of the voice
Shimmer: DDA	Shimmer (dB-by-delta amplitude), a measure of the variation in the amplitude of the voice
NHR	Noise-to-Harmonics Ratio, a measure of the amount of noise in the voice signal
HNR	Harmonics-to-Noise Ratio, a measure of the amount of noise in the voice signal
Status	Health status of the patient, where a 0 = subject is healthy, and 1 = subject has Parkinson's disease
RPDE	Recurrence Period Density Entropy, a measure of the complexity of the voice signal
DFA	Detrended Fluctuation Analysis, a measure of the long-range dependence in the voice signal
Spread1	The first spectral moment, statistical measure of the voice signal's frequency spread
Spread2	The second spectral moment, a statistical measure of the voice signal's frequency spread
D2	A nonlinear dynamic parameter, a statistical measure of the complexity of the voice signal
PPE	Pitch Period Entropy, a measure of the variation in fundamental frequency

3.2. Data Preprocessing and Feature Engineering

Data preprocessing and feature engineering are essential steps for developing reliable and generalizable machine learning (ML) models. Before model training, the dataset was examined for missing values and outliers to ensure data integrity and consistency. Feature scaling was applied to normalize the range of input variables, which is particularly important for distance-based classifiers such as K-Nearest

Neighbors. These preprocessing steps help stabilize model training and reduce the risk of biased predictions.

Class imbalance is a common issue in medical datasets and can negatively impact model performance, particularly for minority-class predictions. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training set to prevent data leakage. SMOTE generates synthetic samples for the minority class by interpolating between existing instances. After applying SMOTE, the training dataset achieved a balanced class distribution with 118 samples per class, while the test set remained untouched to ensure a fair and unbiased evaluation.

In addition to class balancing, correlation-based feature analysis was performed to reduce redundancy and mitigate overfitting. Pearson's correlation coefficient was used to quantify linear relationships between features. Features exhibiting a correlation coefficient greater than 0.8 were considered highly correlated and redundant. Figure 2 illustrates the correlation matrix before feature removal, highlighting several strongly correlated feature pairs. Based on this analysis, redundant features were removed, and the resulting reduced feature set was used for model training. Figure 3 presents the correlation matrix after the removal of highly correlated features, demonstrating a more compact and less redundant feature representation.

This preprocessing pipeline ensures that the models are trained on balanced, non-redundant, and appropriately scaled data, thereby improving model robustness, interpretability, and reproducibility.

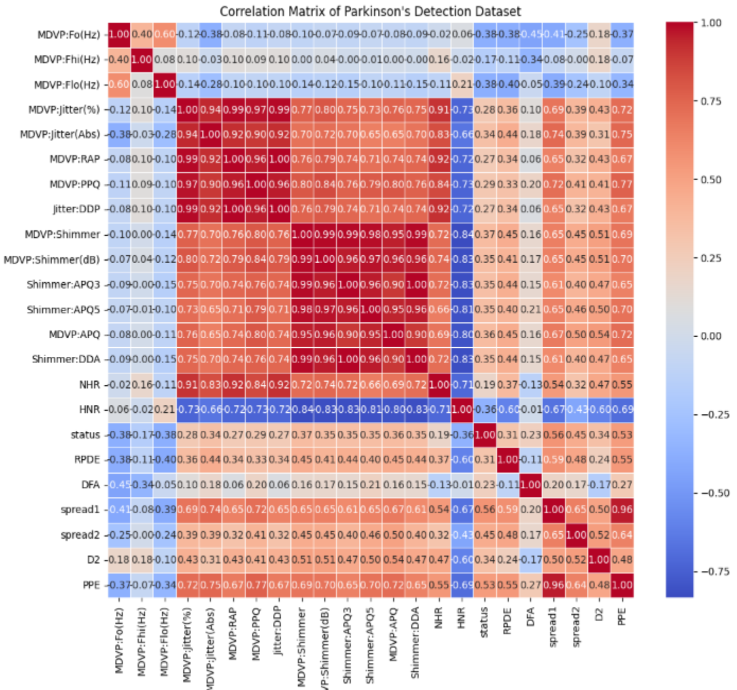


Figure 2. Correlation Matrix before removing highly correlated features

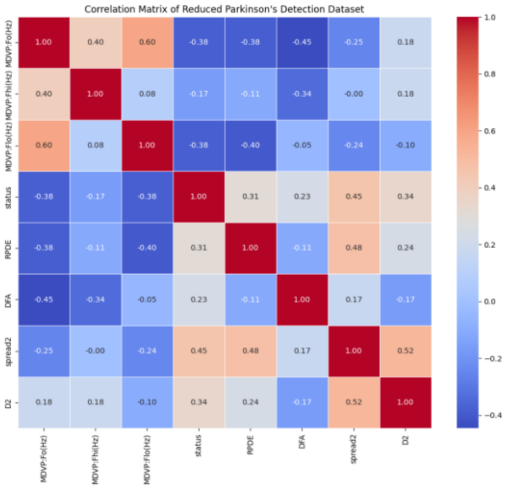


Figure 3. Correlation Matrix after removing highly correlated features

3.3. Model Training

The dataset was partitioned into training and testing subsets using an 80:20 split. All model optimization and resampling procedures were performed exclusively on the training set to prevent data leakage. Hyperparameter tuning was conducted using grid search combined with stratified cross-validation to identify optimal model configurations. Performance optimization was guided by balanced evaluation metrics to ensure reliable classification of both Parkinson’s disease and healthy subjects.

A diverse set of classical machine and deep learning models was evaluated to identify the most effective approach for Parkinson’s disease detection using vocal features.

3.3.1. Logistic Regression (LR)

In LR, a logistic function is used to represent a binary dependent variable. LR is a statistical method that can be used to predict the probability of different experimental outcomes [17]. It offers a simple and understandable way to determine the presence of PD by assuming a linear relationship between the features and the log-odds of the outcome. The mathematical representation of LR models in the Equation (1) the probability P of the binary outcome (PD or not) as a function of input features (X). Where, β_0 the intercept term and β_i are the coefficients for the input features X_i.

$$P(Y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}} \quad (1)$$

3.3.2. Support Vector Machine (SVM)

For classification tasks, a supervised learning approach called SVM is employed. It functions by determining the ideal hyperplane that maximally divides data points belonging to various classifications [18]. SVM uses biological voice measurements to categorize people as either healthy or suffering from PD. SVM can handle non-linear relationships in the data because of the usage of kernels. As a result, a strong model is produced that successfully differentiates across classes using intricate feature patterns. SVM aims to find a hyperplane that best separates the two classes (PD or not) which is state in Equation (2). Where, w is the weight vector and b is the bias term.

$$f(X) = w \cdot X + b \quad (2)$$

3.3.3. Random Forest (RF)

During training, a RF Classifier builds several DTs and produces a class that is the mean of the classes (classification) of the individual trees. It employs biological voice measurements to detect Parkinson's illness and build a strong model that reduces overfitting and boosts accuracy. It improves prediction performance and stability by averaging the outcomes of numerous DTs, which helps it distinguish between people in good health and those who have PD. Equation (3) shown how final prediction is made by aggregating the predictions of all individual trees, typically through majority voting. Where, T_i are the individual DTs in the forest.

$$\hat{Y} = \mathrm{mode}\big(T_1(X), T_2(X), \ldots, T_m(X)\big) \quad (3)$$

3.3.4. Gradient Boosting (GB)

An ensemble learning technique called a GB Classifier develops several weak learners in a sequential manner. The model employs biological voice measurements to iteratively repair prior trees' faults in PD identification. Every new tree is trained using the residuals (errors) of the older trees, with an emphasis on situations that are challenging to categorize. The accuracy and robustness of the model are enhanced by this procedure. Equation (4) represents the final prediction. Where, α_m are the learning rates and h_m are the weak learners.

$$\hat{Y} = \sum_{m=1}^{M} \alpha_m h_m(X) \quad (4)$$

3.3.5. eXtreme Gradient Boosting (XGBoost)

XGBoost is a distributed GB library that has been tuned for maximum efficiency, versatility, and portability. It is a sophisticated version of the GB algorithm that is frequently applied to regression and classification applications in ML [19]. It sequentially builds a set of DTs, with each tree trained to fix the mistakes of its predecessors. Because of its reputation for scalability, speed, and performance, the prediction is similar to Equation (4), with additional regularization terms to prevent overfitting. Where, regularization is applied to control the complexity of the trees.

3.3.6. K-Nearest Neighbors (KNN)

It is a powerful algorithm that is widely used in the field of ML. It is particularly useful for classification tasks, as it allows us to compare a data point with its closest neighbors in order to make accurate predictions [20]. Regarding the Parkinson's dataset, the model is configured with k=3, which implies that it considers the three closest data points for classification. It computes the distance (typically Euclidean) between the test instance and all training instances. The test instance is assigned to the class that has the highest number of neighbors. This approach assumes that data points with similar characteristics (voice measurements and other features) belong to the same category (healthy or Parkinson's). KNN classifies a new instance based on the majority class among its KNN. Equation (5) demonstrate the decision rule. Where, $\mathrm{Y}_{\mathrm{NN}_\mathrm{i}}$ are the classes of the KNN.

$$\widehat{\mathrm{Y}} = \mathrm{mode}\left(\mathrm{Y}_{\mathrm{NN}_1}, \mathrm{Y}_{\mathrm{NN}_2}, \dots, \mathrm{Y}_{\mathrm{NN}_\mathrm{k}}\right) \tag{5}$$

3.3.7. Gaussian Naive Bayes (GNB)

It is a classifier that uses probabilistic calculations and assumptions about feature independence and feature distribution to make predictions. Within the context of the Parkinson's dataset, it computes the likelihood of each category (healthy or Parkinson's) for a given instance by analyzing its characteristics. The model assigns the class with the highest calculated probability to the instance. This approach is highly effective when dealing with small datasets and data with a high number of dimensions. It is especially valuable in this context because of its straightforwardness and effectiveness in managing the different characteristics obtained from voice measurements. GNB calculates the posterior probability using Bayes' theorem (Equations 6 and 7), assuming feature independence and Gaussian distribution. Where, u_{ic} and σ_{ic} are the mean and standard deviation of feature X_i for class c.

$$\mathrm{P}(\mathrm{Y} = \mathrm{c}|\mathrm{X}) = \frac{\mathrm{P}(\mathrm{Y}=\mathrm{c}) \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{P}(\mathrm{X}_\mathrm{i}|\mathrm{Y} = \mathrm{c})}{\mathrm{P}(\mathrm{X})} \tag{6}$$

$$\mathrm{P}(\mathrm{X}_\mathrm{i}|\mathrm{Y} = \mathrm{c}) = \frac{1}{\sqrt{2\pi\sigma_{\mathrm{ic}}^2}} \exp\left(-\frac{(\mathrm{X}_\mathrm{i} - \mu_{\mathrm{ic}})^2}{2\sigma_{\mathrm{ic}}^2}\right) \tag{7}$$

3.3.8. Decision Tree (DT)

Decision Tree (DT) constructs hierarchical decision rules by recursively partitioning the feature space using impurity-based criteria such as Gini index or entropy. Each internal node represents a decision based on a feature threshold, and each leaf node corresponds to a class label. While highly interpretable, decision trees are susceptible to overfitting and are therefore included primarily as comparative baselines.

3.3.9. AdaBoost

AdaBoost is an ensemble learning method that iteratively combines multiple weak learners, typically decision trees, by assigning higher weights to misclassified samples in successive iterations. The final prediction is obtained through a weighted majority vote of the individual learners.

3.3.10. 1D Convolutional Neural Network (1D CNN)

The 1D CNN is a method that is intended to isolate local features in sequential or time-series data. This research, it deals with sequences of vocal features to determine useful patterns related to Parkinson's disease. Convolutional layers store time-related dependencies, and pooling layers decrease the dimensionality and enhance generalization, and then, fully connected layers are used to perform classification.

3.3.11. 2D Convolutional Neural Network (2D CNN)

The 2D CNN is used to process two-dimensional data of the vocal data, including spectrograms or feature matrices. Convolutional layers learn spatial features and hierarchical features presentation, enabling the model to acquire intricate relationships among features. Layers with pooling decrease the spatial dimensions, and the last layers of the network are fully connected and do the ultimate classification.

3.3.12. Long Short-Term Memory Network (LSTM)

LSTM is a kind of recurrent neural network that is specially created to represent long-term reliance in sequential information. It has been useful in identifying the temporal patterns in the sequences of vocal features to detect Parkinson's disease. The LSTM units make use of memory cells and gating systems to selectively remember and forget, and therefore, the network can learn short-term and long-term trends in the data.

3.4. Model Interpretation and Explanation

Feature importance and explainable AI methods are essential for interpreting ML models. Feature importance quantifies each input feature's contribution to a model's predictions, helping identify crucial features in the decision-making process. Permutation importance, a specific method of feature importance, assesses the impact of randomly rearranging feature values on model performance, revealing the model's dependency on that feature. Mathematically, if $(\hat{\mathrm{f}})$ is the trained model and $\left(\mathrm{L}(\hat{\mathrm{f}})\right)$ is the loss function, the permutation importance of feature (j) is $\mathrm{Importance(j)} = \mathrm{L}(\hat{\mathrm{f}}) - \mathrm{L}\left(\widehat{\mathrm{f}_{\pi(j)}}\right)$. Where $\left(\widehat{\mathrm{f}_{\pi(j)}}\right)$ represents the model's prediction after shuffling feature (j).

LIME is another powerful method for explaining individual predictions by approximating the model locally with an interpretable model, such as a linear model. LIME achieves this by generating perturbed samples around the instance of interest and fitting a simple, interpretable model, such as $(\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)$,, where $(\hat{y})$ is the predicted value, (β_i) are the coefficients, and $(\mathrm{x_i})$ are the feature values.

3.5. Evaluation Metrics

For binary classification, we assessed the performance of each model using standard evaluation metrics including accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the classification, while precision quantifies the proportion of true positive predictions among all positive predictions. Recall, also known as sensitivity, measures the proportion of true positives correctly identified by the model. The F1-score, the harmonic mean of precision and recall, provides a balanced measure of the model's performance. These metrics can be computed using Equation 10-13.

$$\mathrm{Accuracy} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{10}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{11}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{12}$$

$$\mathrm{F1} = \frac{2\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision+Recall}} \tag{13}$$

4. Experimental Setup and Result Analysis

4.1. Comparative Analysis and Visualization of Machine Learning Models

Figure 4 depicts the performance of ML models across four metrics: Accuracy, Precision, Recall, and F1 Score. Among these models, KNN emerges as the best performer, consistently achieving high scores across all metrics. KNN demonstrates an impressive balance with its Accuracy closely matching that of advanced ensemble methods like RF, GB, XGBoost, and AdaBoost. Additionally, KNN maintains high Precision, Recall, and F1 Score, highlighting its robustness and reliability in classification tasks. While LR, SVM, and the ensemble methods also exhibit strong performance, KNN's consistency across all evaluation criteria sets it apart. In contrast, Naive Bayes and DT show relatively moderate performance, particularly with lower Recall and F1 Score.

The learning curve for the KNN model (Figure 6) depicts the relationship between the number of training examples and the model's performance, including training score and cross-validation score. The red curve reflects the training score, demonstrating strong performance on training data, while the green curve illustrates the cross-validation score, reflecting model performance on unseen data. Initially, the training score is high, indicating a good fit to the training data, while the cross-validation score is lower, suggesting poor generalization. As the number of training examples increases, both scores improve, indicating reduced overfitting and better generalization. The convergence of the two scores suggests improved model performance with more data. The shaded regions around the curves represent variability, with narrower regions indicating lower variability. The confusion matrix of KNN is depicted in Figure 5.

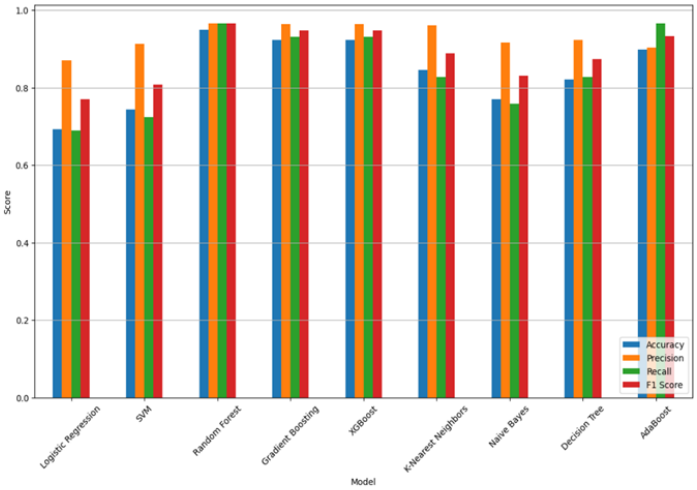


Figure 4. Comparison of Model Performance

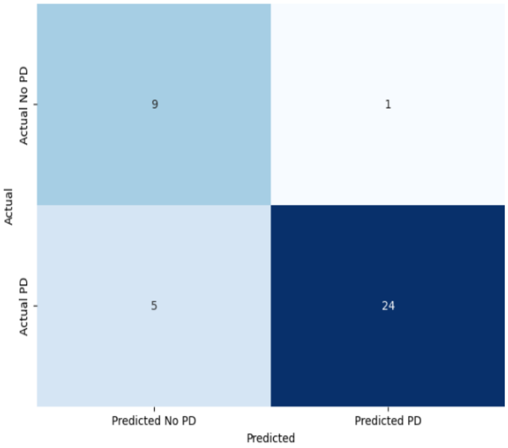


Figure 5. Confusion Matrix for KNN

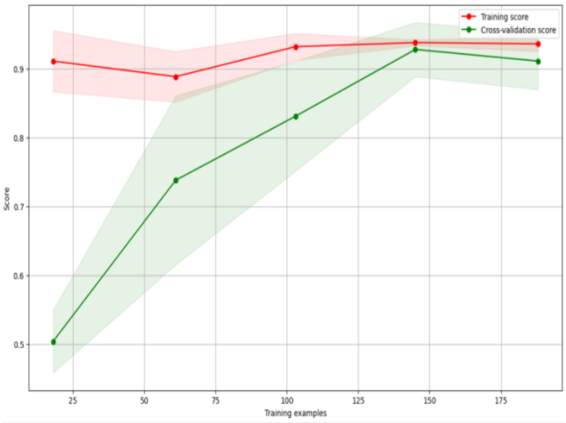


Figure 6. Learning Curve for KNN.

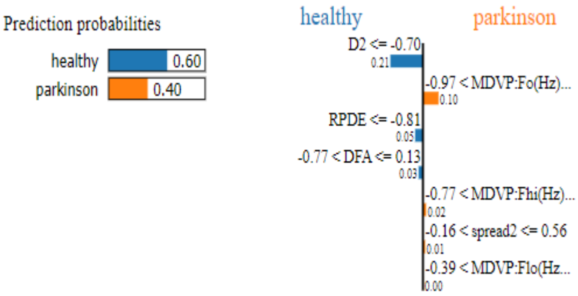
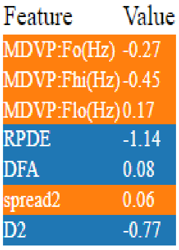


Figure 7. Prediction probabilities of KNN

Figure 7 displays prediction probabilities for two classes: ‘healthy’ (0.60) and ‘parkinson’ (0.40). On the right side, features with corresponding weights are listed. Notably, “D2 <= -0.70” has the highest weight (around 0.30), favoring a 'Parkinson' prediction, while other features contribute differently. In Figure 8, features with positive values are considered to have a positive impact on the model's outcome, while those with negative values have a lesser or negative impact. For instance, 'MDVP:Flo(Hz)' has a positive value of 0.17, indicating a favorable influence, whereas 'RPDE' has a negative value of -1.14, suggesting a less significant or negative influence.



Feature	Value
MDVP:Fo(Hz)	-0.27
MDVP:Fhi(Hz)	-0.45
MDVP:Flo(Hz)	0.17
RPDE	-1.14
DFA	0.08
spread2	0.06
D2	-0.77

Figure 8. Feature analysis of the KNN model

The study into vocal biomarkers for the early identification and monitoring of PD using ML models represents a notable breakthrough in medical diagnostics. Through the analysis of voice samples, the study utilizes a non-intrusive and easily accessible data source, potentially enabling more convenient and timely diagnosis for patients. It is worth mentioning the significance of explainable AI, as it not only assists in making predictions but also offers transparency and comprehension of how the models make decisions. Recognizing Spread2, RPDE, and MDVP (Hz) as significant vocal biomarkers highlights the complex connection between voice and neurological conditions. Spread2's analysis of pitch fluctuation, RPDE's assessment of voice pattern complexity, and MDVP's evaluation of fundamental frequency range provide a comprehensive method for comprehending the effects of PD on speech. Observing the lower fundamental frequencies in Parkinson's patients, as indicated by MDVP (Hz), could potentially be a valuable indicator for early detection, prompting additional medical investigation. The study's alignment with previous research not only confirms these findings but also adds to the existing body of knowledge, contributing to the ongoing efforts to combat PD. Considering the future, the proposal to include time-series voice data and a wider range of demographics in upcoming research demonstrates a dedication to improving the strength and understandability of the models. Figures 7 and 8 highlights both the KNN model's prediction probabilities for 'healthy' and 'Parkinson' classes and the relative impact of individual vocal features, with Spread2, RPDE, and MDVP (Hz) identified as key biomarkers influencing the model's decisions. With these advancements, healthcare solutions can become more personalized and accurate, resulting in a significant improvement in the quality of life for individuals impacted by PD. The combination of ML and vocal analysis in medical research shows great potential, and the results of this study have the potential to inspire new and creative approaches in the healthcare industry.

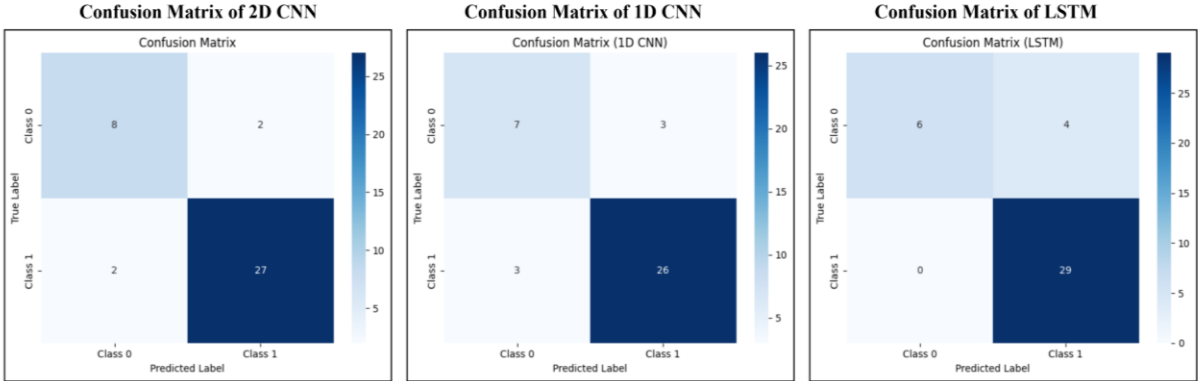
4.2. Comparative Analysis and Visualization of Deep Learning Models

Table 2 presents the results of three deep learning models: 1D CNN, LSTM, and 2D CNN in the classification of Parkinson's disease. The test accuracy of LSTM and 2D CNN was highest with 89.74% and 84.62%, respectively, whereas 1D CNN was 84.62. The fact that the training and test accuracies are similar indicates that, once the models are trained, they all generalize well to unseen data. The reason why LSTM works well is probably that it is sensitive to time trends in the data, whereas 2d CNN is sensitive to spiciness on the image representations. All in all, LSTM and 2D CNN are more suitable for accurately diagnosing Parkinson's disease. Confusion matrices are presented in Figure 9 for all deep learning models to provide a detailed view of their classification behavior.

Table 2. Training and test accuracies of different deep learning models for Parkinson's Disease diagnosis.

Model	Training Accuracy	Test Accuracy
1D CNN	0.8526	0.8462
LSTM	0.8718	0.8974
2D CNN	0.8718	0.8974

Confusion matrices of 1D CNN, 2D CNN, and LSTM models on the test set are provided in Figure 9. The 1D CNN with 3 false negatives and 3 false positives rightly classified 26 samples of Class 1 and 7 of Class 0. The 2D CNN with 2 false negatives and 2 false positives, and rightly classified 27 samples of Class 1 and 8 of Class 0. The LSTM model with 0 false positives and 4 false negatives rightly classified 29 samples of Class 1 and 6 of Class 0.


Confusion Matrix of 2D CNN

Confusion Matrix of 1D CNN

Confusion Matrix of LSTM

Figure 9. Confusion matrix of all evaluated deep learning models.

5. Conclusion

These results indicate the potential of voice-based analysis in accurately diagnosing Parkinson's disease, with both machine learning and deep learning models able to successfully differentiate between patients with the condition and without it, and deep learning models (1D CNN, 2D CNN, and LSTM) show better predictive accuracy. Rigorous data handling and model selection guarantee the credibility of the tools, and by tackling the issue of class imbalance, deep learning models can be improved to achieve better predictive performance. Methods such as LIME and permutation importance provide valuable insights into model predictions, assisting clinicians in making informed decisions. Our study is limited by the use of a single, canonical voice dataset, which may not fully capture the variability in real-world clinical settings. Additionally, the model's performance on multi-modal or larger, more complex datasets remains untested and requires further investigation. Efforts are continuously being made to improve the models' strength and ability to apply to various situations. This is achieved through long-term research, validation using different datasets, and working closely with medical professionals. With the incorporation of diagnostic techniques based on ML, there is great potential for early detection and personalized management of PD. This can ultimately result in better patient outcomes and improved quality of life.

CRediT Author Contribution Statement

Ferdaus Ibne Aziz: Conceptualization, Methodology, Software, Data curation, Formal analysis, Writing – original draft; Daniel Ojeda Rosales: Conceptualization, Methodology, Validation, Writing – original draft, Becky Firomssa Gudeta: Validation, Resources, Visualization, Writing – original draft; Corresponding author; Jia Uddin: Supervision, Fund Acquisition, Writing – review & editing;

Acknowledgement

This research is funded by Woosong University Academic Research 2026.

References

[1] Divya M. Radhakrishnan and Vinay Goyal, "Parkinson's disease: A review", *Neurology India*, Print ISSN 0028-3886, Online ISSN 1998-4022, March 2018, Vol. 66, No. 7, pp. S26–S35, Published by Medknow Publications (publication of the Neurological Society of India), DOI: 10.4103/0028-3886.226451, Available: https://journals.lww.com/neur/fulltext/2018/66001/parkinson_s_disease__a_review.7.aspx.

[2] Jie Mei, Christian Desrosiers and Johannes Frasnelli, "Machine Learning for the Diagnosis of Parkinson's Disease: A Review of Literature", *Frontiers in Aging Neuroscience*, Online ISSN: 1663-4365, 6 May 2021, Vol. 13, pp. 1-41,

Published by Frontiers, DOI: 10.3389/FNAGI.2021.633752/BIBTEX, Available: https://www.frontiersin.org/journals/aging-neuroscience/articles/10.3389/fnagi.2021.633752/full.

[3] Zehra Karapinar Senturk, "Early diagnosis of Parkinson's disease using machine learning algorithms", *Medical Hypotheses*, Print ISSN: 0306-9877, May 2020, Vol. 138, pp. 1-5, DOI: 10.1016/J.MEHY.2020.109603, Available: https://pubmed.ncbi.nlm.nih.gov/32028195/.

[4] Gunjan Pahuja and T. N. Nagabhushan, "A Comparative Study of Existing Machine Learning Approaches for Parkinson's Disease Detection", *IETE Journal of Research*, Print ISSN: 0377-2063, Online ISSN: 0974-780X, 22 October 2018, Vol. 67, No. 1, pp. 4–14, Published by Taylor & Francis Ltd, DOI: 10.1080/03772063.2018.1531730, Available: https://www.tandfonline.com/doi/abs/10.1080/03772063.2018.1531730.

[5] Oliveira M Oliveira, Luis Coelho, Eduardo Carvalho, Manuel J Ferreira-Pinto, Rui Vaz *et al.*, "Machine learning for adaptive deep brain stimulation in Parkinson's disease: closing the loop", *Journal of Neurology*, Print ISSN: 0340-5354, Online ISSN: 1432-1459, 02 August 2023, Vol. 270, No. 11, pp. 5313-5326, Published by Springer, DOI: 10.1007/s00415-023-11873-1, Available: https://link.springer.com/article/10.1007/s00415-023-11873-1.

[6] Bruno Fonseca Oliveira Coelho, Ana Beatriz Rodrigues Massaranduba, Carolline Angela dos Santos Souza, Giovanni Guimarães Viana, Ivani Brys *et al.*, "Parkinson's disease effective biomarkers based on Hjorth features improved by machine learning", *Expert Systems with Applications*, Print ISSN: 0957-4174, Online ISSN: 1873-6793, February 2023, Vol. 212, pp. 1-9, Published by Elsevier, DOI: 10.1016/j.eswa.2022.118772, Available: https://www.sciencedirect.com/science/article/abs/pii/S0957417422017900.

[7] Aditi Govindu and Sushila Palwe, "Early detection of Parkinson's disease using machine learning", *Procedia Computer Science*, Online ISSN: 1877-05092023, 2023, Vol. 218, pp. 249–261, Published by Elsevier, DOI: 10.1016/j.procs.2023.01.007, Available: https://www.sciencedirect.com/science/article/pii/S1877050923000078.

[8] Max Little, Patrick McSharry, Eric Hunter, Jennifer Spielman and Lorraine Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease", *IEEE Transactions on Biomedical Engineering*, Print ISSN: 0018-9294, Online ISSN: 1558-2531, 30 September 2008, Vol. 56, No. 4, pp. 1015-1022, Published by Institute of Electrical and Electronics Engineers (IEEE), DOI: 10.1109/TBME.2008.2005954, Available: https://ieeexplore.ieee.org/document/4636708.

[9] Liaqat Ali, Ashir Javeed, Adeeb Noor, Hafiz Tayyab Rauf, Seifedine Kadry *et al.*, "Parkinson's disease detection based on features refinement through L1 regularized SVM and deep neural network", *Scientific Reports*, Online ISSN: 2045-2322, January 2024, Vol. 14, No. 1, pp. 1–14, Published by Nature Portfolio, DOI: 10.1038/s41598-024-51600-y, Available: https://www.nature.com/articles/s41598-024-51600-y.

[10] Osmar Pinto Neto, "Harnessing Voice Analysis and Machine Learning for Early Diagnosis of Parkinson's Disease: A Comparative Study Across Three Datasets", *Journal of Voice*, Online ISSN: 1873-4588, 12 May 2024, Article in Press, pp. 1-7, Published by Elsevier, DOI: 10.1016/J.JVOICE.2024.04.020, Available: https://www.sciencedirect.com/science/article/abs/pii/S0892199724001395.

[11] Lizbeth Naranjo, Carlos J. Perez, Jacinto Martin and Yolanda Campos-Roca, "A two-stage variable selection and classification approach for Parkinson's disease detection by using voice recording replications", *Computer Methods and Programs in Biomedicine*, Online ISSN: 0169-2607, April 2017, Vol. 142, pp. 147–156, Published by Elsevier, DOI: 10.1016/J.CMPB.2017.02.019, Available: https://www.sciencedirect.com/science/article/abs/pii/S0169260716302206.

[12] Gunjan Pahuja and T. N. Nagabhushan, "A Comparative Study of Existing Machine Learning Approaches for Parkinson's Disease Detection", *IETE Journal of Research*, Print ISSN: 0377-2063, Online ISSN: 0974-780X, 22 October 2018, Vol. 67, No. 1, pp. 4-14, Published by Taylor & Francis Ltd, DOI: 10.1080/03772063.2018.1531730, Available: https://www.tandfonline.com/doi/abs/10.1080/03772063.2018.1531730.

[13] Liaqat Ali, Ce Zhu, Zhonghao Zhang and Yipeng Liu, "Automated Detection of Parkinson's Disease Based on Multiple Types of Sustained Phonations Using Linear Discriminant Analysis and Genetically Optimized Neural Network", *IEEE Journal of Translational Engineering in Health and Medicine*, Online ISSN: 2168-2372, 7 October 2019, Vol. 7, pp. 1-10, Published by Institute of Electrical and Electronics Engineers (IEEE), DOI: 10.1109/JTEHM.2019.2940900, Available: https://ieeexplore.ieee.org/document/8861144.

[14] Wu Wang, Junho Lee, Fouzi Harro, and Ying Sun "Early Detection of Parkinson's Disease Using Deep Learning and Machine Learning", *IEEE Access*, Online ISSN: 2169-3536, 12 August 2020, Vol. 8, pp. 147635–147646, Published by Institute of Electrical and Electronics Engineers (IEEE), DOI: 10.1109/ACCESS.2020.3016062, Available: https://ieeexplore.ieee.org/document/9165732.

[15] Hakan Gunduz, "Deep Learning-Based Parkinson's Disease Classification Using Vocal Feature Sets", *IEEE Access*, Online ISSN: 2169-3536, 20 August 2019, Vol. 7, pp. 115540-115551, Published by Institute of Electrical and Electronics Engineers (IEEE), DOI: 10.1109/ACCESS.2019.2936564, Available: https://ieeexplore.ieee.org/document/8807125.

[16] Rizgar Zebari, Adnan Abdulazeez, Diyar Zeebaree, Dilovan Zebari and Jwan Saeed, "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction", *Journal of Applied Science and Technology Trends*, ISSN: 2708-0757, 15 May 2020, Vol. 1, No. 1, pp. 56–70, Published by Interdisciplinary Publishing Academia, DOI: 10.38094/jastt1224, Available: https://jastt.org/index.php/jasttpath/article/view/24.

[17] Rezaul Haque, Md Babul Islam, B. D. Parameshachari, Katura Gania Khushbu, Shafiur Rahman *et al.*, "Bengali Emotion Classification Using Hybrid Deep Neural Network", in *Proceedings of the 2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE)*, 2-3 November 2023, Ballari, India, pp. 1-7, DOI: 10.1109/AIKIIE60097.2023.10389834, Available: https://ieeexplore.ieee.org/document/10389834.

[18] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines", *IEEE Intelligent Systems and their Applications*, Print ISSN: 1094-7167, 31 August 1998, Vol. 13, No. 4, pp. 18-28, Published by Institute of Electrical and Electronics Engineers (IEEE), DOI: 10.1109/5254.708428, Available: https://ieeexplore.ieee.org/document/708428.

[19] Kashif Shaheed, Qaisar Abbas, Ayyaz Hussain and Imran Qureshi, "Optimized Xception Learning Model and XgBoost Classifier for Detection of Multiclass Chest Disease from X-ray Images", *Diagnostics*, ISSN: 2075-4418, 3 August 2023, Vol. 13, No. 15, pp. 2583, Published by Multidisciplinary Digital Publishing Institute (MDPI), DOI: 10.3390/DIAGNOSTICS13152583, Available: https://www.mdpi.com/2075-4418/13/15/2583.

[20] Mustafa Çakir, Mesut Yilmaz, Mükerrem Atalay Oral, Hüseyin Özgür Kazanci and Okan Oral, "Accuracy assessment of RFerns, NB, SVM, and kNN machine learning classifiers in aquaculture", *Journal of King Saud University - Science*, Print ISSN: 1018-3647, Online ISSN: 2213-686X, 21 June 2023, Vol. 35, No. 6, Article No. 102754, Published by Elsevier B. V., DOI: 10.1016/J.JKSUS.2023.102754, Available: https://www.sciencedirect.com/science/article/pii/S1018364723002161.