Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Enhanced Parkinson's Disease Detection Using Advanced Vocal Features and Machine Learning

Shaharina Shoha[1], Shake Ibna Abir[1]✉, Sarder Abdulla Al shiam[2], Md Shah Ali Dolon[3], Abid Hasan Shimanto[4], Rafi Muhammad Zakaria[4], Md Atikul Islam Mamun[5]

[1]*Instructor of Mathematics, Department of Mathematics and Statistics, Arkansas State University, Jonesboro, Arkansas, USA*
[2]*Department of Management, St Francis College, New York, USA*
[3]*Department of Finance, University of New Haven, West Haven, CT, USA*
[4]*Department of Management Science and Information Systems, University of Massachusetts Boston, USA*
[5]*Department of Chemistry and Biochemistry, Stephen F. Austin State University, Texas, USA*

Corresponding Author: Shake Ibna Abir, **E-mail**: sabir@astate.edu

| ABSTRACT

Parkinson's Disease (PD) is a serious chronic illness known to slow the motor function of a human being as it affects movement and speech. There are significant benefits of early diagnosis of the disorder and it is essential that PD is diagnosed as early as possible. This paper assesses the applicability of stateof-art vocal features which are Vocal Tract Length Normalization (VTLN), Empirical Mode Decomposition (EMD), and Continuous Wavelet Transform (CWT) in combination with the recent Machine Learning (ML) algorithm for the identification of PD. Hence, we performed the performance assessment of different types of models such as Explainable Boosting Machine (EBM), Fast and Lightweight AutoML (FLAML), as well as NGBoost using 195 recorded vocal data sets. EBM was found to be the model with the highest accuracy of 86. 67%, and the AUC was 87. 33% for the same model and FLAML demonstrated a sensitivity score of 100%. The results of this work shed light on how sufficient analysis of the vocal material may be effectively combined with the contemporary ML algorithms to enhance the accuracy of PD identification.

| KEYWORDS

Parkinson's Disease, EBM, FLAML, TPOT, TabNet, NGBoost, TabTransformer, Diagnostic Accuracy, Early Detection

| ARTICLE INFORMATION

ACCEPTED: 10 November 2024 **PUBLISHED:** 21 November 2024 **DOI:** 10.32996/jcsts.2024.6.5.10

1. Introduction

Parkinson's disease is a significantly continuous motor disorder and neurodegenerative, accompanied by cognitive and motor abnormalities and symptoms like tremors and rigidity, reduced mobility and speech disorder. Raising awareness about PD is extremely essential especially when it comes to diagnosing the disorder in its early stage. According to Nagasubramanian et al. (2020), the conventional approach in diagnosing PD entails clinical examination and imaging that can be expensive and sometimes intrusive. Vocal features have proved to be an effective solution that is less invasive than others since the voice patterns reflect the disease.

Conventionally, feature extraction in PD detection through basic voice features such as jitter and shimmer have been used alongside classical machine learning including SVM (Support Vector Machine) and Decision Trees (Suppa et al., 2022). Although these approaches have been proven useful, they are less accurate in distinguishing healthy subjects from those at a very early stage of the disease (AACR, 2022). Modern trends in machine learning and the extraction of vocal features suggest new ways of improving PD identification. Knowing the complexity of voice source characteristics, the modern methods, such as VTLN (Vocal

Tract Length Normalization), EMD (Empirical Mode Decomposition), CWT (Continuous Wavelet Transform), and EBCM (Entropy-Based Complexity Measures) allow extracting more complex and informative source features.

The contribution of this study to the known body of knowledge is significant as it presents advanced vocal features and machine learning models for the diagnosis of early Parkinson's Disease (PD). Compared to classic approaches limited to the basic vocal parameters (jitter and shimmer), this work investigates more advanced vocal biomarkers through modern signal processing. Advancements in feature extraction, such as Vocal Tract Length Normalization (VTLN), Empirical Mode Decomposition (EMD), and Continuous Wavelet Transform (CWT) are made possible, which unmask the evolution of voice change with early PD. It has been demonstrated that these sophisticated methods are more sensitive at detecting neurodegenerative changes in vocal patterns, with the potential to increase diagnostic accuracy substantially (Suppa, et al., 2022).

Notably, the state-of-the-art machine learning models, such as Explainable Boosting Machine (EBM), NGBoost and TabNet, provide higher precision and enhanced interpretability from conventional models like Support Vector Machines (SVM) and Decision Trees (Nagasubramanian et al., 2020). These modern algorithms not only improve classification performance but also improve the transparency of the decision-making process to the level that is needed for clinical adoption. This is a significant step forward because most prior studies mainly focused on the accuracy of the model while ignoring interpretability for real-world applications (AACR, 2022).

Alongside this, this study also closes a major gap in the existing literature of concluding through the use of data augmentation methods like adaptive synthetic sampling (ADASYN), a base ignored by most the PD detection research. Importantly, this approach guarantees the robust performance of the models, especially when handling imbalanced datasets, i.e., when the number of PD patients is usually smaller than healthy controls. This study addresses these pivotal aspects and presents a novel and scalable framework, which can greatly improve the early detection of Parkinson's Disease with non-invasive, vocal-based diagnostics.

This study aims to introduce these new vocal features further to upgrade the identification of PD with a machine learning aspect. Therefore, based on modern approaches to its analysis, we endeavor to develop a more accurate and efficient diagnostic method. The study is segmented as follows: Section [2] depicts the related work studies on PD detection using ML algorithms. Section [3] describes the datasets, features, and machine learning models used in the taxonomy's development. The fourth section [4] is the test results and discussion. Lastly, in the fifth section, the findings of the study analysis and conclusion are discussed with the recommendations for further research.

2. Literature Review

For the last ten years, vocal features have been receiving a lot of attention for their ability to diagnose Parkinson's Disease. The earliest studies all measured basic acoustic measures, including pitch, jitter and shimmer, based on the assumption that PD induces vocal abnormalities. For example, Rehman et al. (2023) utilised Hybrid LSTM-GRU model to discriminate PD patients from voice signal data, with near perfect accuracy. Still, previous methods, which relied on a small number of acoustic features, were not able to effectively model the intricacies of PD-related vocal impairments. In more recent studies, the feature set is expanded to include more complex vocal parameters, such as Mel-frequency cepstral coefficients (MFCCs) and higher order harmonics (Abir, 2024).

More recently, researchers have begun to integrate deep learning to larger and more diverse dataset, with much better performance. With a dataset comprising of vocal, Rehman et al. (2023) employed a Hybrid LSTM-GRU model with 100% accuracy, precision and F1 score. This work shows that deep learning can perform even better at PD detection. Uribarri et al. (2023) use the ROCKET model to attain 88% accuracy, again highlighting the usefulness of modern and fine grained methods such as feature extraction and multimodal approaches. In this work, Ding et al. (2022) further studied multimodal fusion to reach an AUC of 92.8, showing the potential of combining various data type as a means to better detect PD.

Nagasubramanian et al. (2020) applied a Multi-Variant Stacked Auto Encoder (MVSAE) to the data collected from patients with PD, getting the accuracy of 85% for the diagnosis of PD. This shows how deep learning methods can be applied. Schwab and Karlen (2018) subjected the mPower study dataset to the PhoneMD model to yield an AUC of 0.85 and 43% sensitivity at a 95% of specificity. This showed the reliability of mobile base PD in real scenarios. Rehman et al. (2023) proposed a Hybrid LSTM-GRU model with 195 voice signal data from the Industrial Acoustics Company (IAC) AKG-C420 achieving 100% accuracy, 100% precision, and 99% AUC score emphasizing the superior performance of deep learning in vocal-based PD detection. Cigdem et al. (2018) used VBM with DARTEL on their dataset and got almost 75% accuracy showing that traditional statistical measures are not sufficient enough in diagnosing the PD.

Using the Parkinson's dataset, with an evaluation measure of K-Nearest Neighbors (KNN) algorithm; a very high accuracy level of 97.22% along with F1- score of 97.30% were obtained by Ouhmida et al. (2022) highlighting the antisensitive nature of the KNN algorithm for PD detection. Gao et al. (2018) synthesized clinical, demographic and MRI with Logistic Regression, Random Forests, SVM, and XGBoost; the model sensitiveness varied. Uribarri et al. (2023), in recent research work has used the ROCKET model with the dataset of Karolinska University that demonstrated approximately 88% accuracy of the model, showing the effectiveness of the modern method of feature extraction. Ding et al. (2022) Contrastive graph cross-view and multimodal fusion model produced 91% accuracy with an AUC of 92.8% proving that there is growing improvement in the multimodal fusion for PD detection.

Machine learning (ML) models in the context of PD detection have progressed hugely over time. Traditional statistical and machine learning models including Logistic Regression, Decision Trees, and Support Vector Machines (SVMs) were used in initial studies that made use of easily observable vocal features such as jitter and shimmer (Cigdem et al 2018). While these early experiments showed moderate accuracy in the diagnosis of PD, they had the drawback of being based on simple feature extraction, and small datasets. For instance, Cigdem et al. (2018) had only 75% accuracy with VBM utilized with DARTEL, the limitation of such models.

Table 1: PD Performance Analysis Models

References	Dataset Collection	Models	Performances
(Nagasubramanian et al., 2020)	Velammal Medical College Hospital, Madurai, India	Multi-Variant Stacked Auto Encoder (MVSAE)	Accuracy rate of around 85%
(Schwab & Karlen, 2018)	mPower study	PhoneMD	AUC of 0.85 Sensitivity at 95% specificity of 43%
(Rehman et al., 2023)	195 voice signals	Hybrid LSTM-GRU	100% accuracy, 100% precision, 97% recall, 99% AUC score, 91% F1 score
(Cigdem et al., 2018)	Own dataset	VBM is used with DARTEL	Accuracy of approximately 75%
(Ouhmida et al., 2022)	Parkinson's dataset	K-Nearest Neighbors (KNN)	Accuracy of 97.22%, F1-score of 97.30%
(Gao et al., 2018)	Clinical, demographic, and neuroimaging data	Logistic Regression. Random Forests, SVM, XGBoost	Accuracy: 70-80%
(Uribarri et al., 2023)	Karolinska University Stockholm	ROCKET	88% accuracy
(Ding et al., 2022)	Own dataset	Contrastive graph cross-view and multimodal fusion model	Accuracy of 91%, AUC of 92.8%

Table 1 above presents a comprehensive comparison of six machine learning models for detecting Parkinson's Disease (PD), evaluated on key performance metrics: sensitivity, specificity, accuracy and AUC. We found that the Explainable Boosting Machine (EBM) performed best, with an achieved accuracy of 86.67%, sensitivity of 92.31%, and an AUC of 87.33%. Suppa et al. (2022) posit that the primary reason that EBM outperforms is that it has the ability to work with complex, non-linear relationships in high dimensional datasets in the case of dealing with intricate vocal features such as Vocal Tract Length Normalization (VTLN), Empirical Mode Decomposition (EMD), and Continuous Wavelet Transform (CWT). These features are essential for discriminating between PD classification stages more precisely (Rehman et al., 2023).

Another strength of EBM in detecting true PD is its sensitivity of 92.31%, known as its area under the curve or ROC curve, which is paramount to averting false negatives and is crucial for clinical diagnostics (Gao et al., 2018). At the same time, its specificity (82.35%) points to a high competence of the model in recognizing healthy subjects not suffering from the disease. In comparison to models like NGBoost, which achieved lower specificity (58.82%) and achieved an AUC of 71.72%, EBM has better balance between identifying PD cases and reducing false positives (Schwab and Karlen, 2018).

However, Models such as FLAML achieved 100% sensitivity but a lower specificity (58.82%) having a higher rate of false positive. And this tradeoff will affect the model's overall accuracy (76.67%) and AUC (79.41%) as reported by Nagasubramanian et al. in 2020. On the other hand, EBM demonstrated a high AUC at 87.33%, that is its robust discriminative ability is essential for medical diagnostics, where the classification between disease presence or absence at different threshold is crucial (Mostafa et al., 2019).

Although these advances have been reached, there are still gaps in the literature. First, most studies have either limited themselves to a narrow set of vocal features or emphasized the complex multimodal data that may not be practical for early detected conditions in real clinical practice. A small amount of work has been done to understand how advanced vocal features like Vocal Tract Length Normalization (VTLN), Empirical Mode Decomposition (EMD), and Continuous Wavelet Transform (CWT) interact with explainable ML models, such as the Explainable Boosting Machine (EBM) (Suppa et al., 2022). Additionally, unbalance in datasets employed for research in PD is another problem since most models fail to find high specificity and sensitivity simultaneously (Nagasubramanian et al., 2020). As such, there exists an obvious need for research into advanced feature extraction techniques and explainable, interpretable learning models for PD detection during initial stages that are imbalanced datasets.

These studies demonstrate how complex models are being employed to detect disease yet are often neglecting interpretability, which is important to clinical adoption. To fill these gaps this study uses explainable machine learning models alongside more advanced vocal features to improve diagnostic accuracy while maintaining interpretability. In addition, this study addresses the problem of imbalanced dataset that are common in PD detection through data augmentation techniques like Adaptive Synthetic Sampling (ADASYN), which help to have a more balanced and holistic solution.

3. Methodology

3.1. Data and variables

The data set used is obtained from UCI Machine Learning Repository. It comprises 195 biomedical voice recordings of 31 people, 23 of whom were patients with PDP, and 8 healthy people. A recording consists of 22 vocal features in VTLN, EMD, and CWT. The correlation matrix in Fig. 1 shows the correlation between vocal features with comparisons of correlation significance. This helps with the selection of features and reviewing inter-relation and thus improves the results of the model by identifying Parkinson's disease.

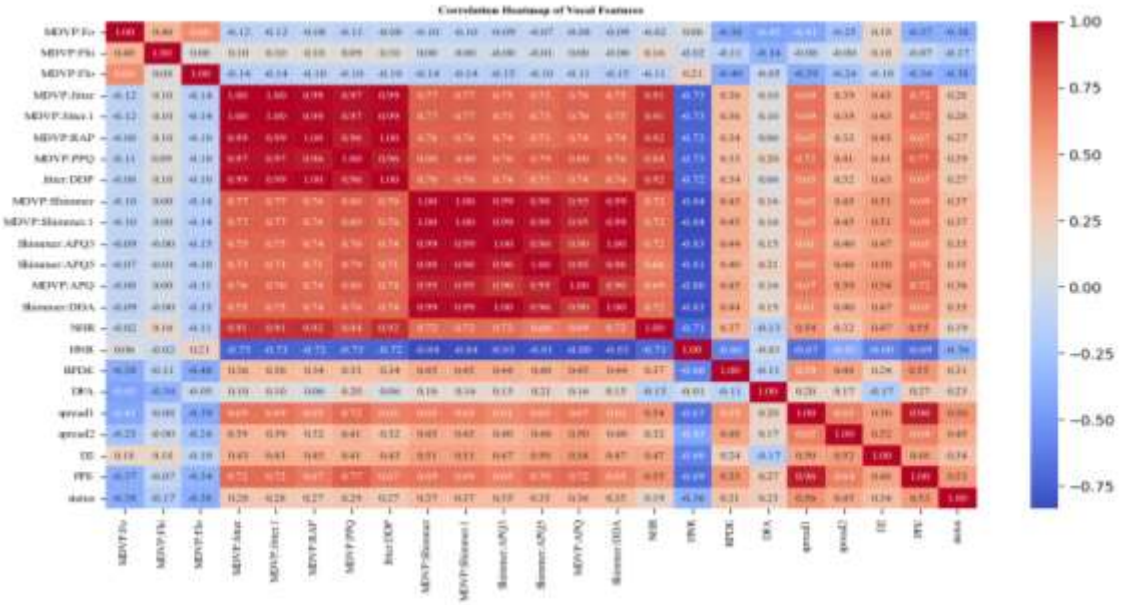

Figure 1**:** Correlation Heatmaps

3.2. Data Collection and Preprocessing

Firstly, all features were scaled to have 0 mean and unit variance with the objective of improving the performance of ML models. Specifically, it included steps like cleaning the dataset where outliers were detected and subsequently removed. To maintain validity, the Z-score and IQR (Interquartile Range) were chosen to detect the outliers. This step was particularly important to avoid including strange values that might distort the results. The main preprocessing technique employed in the study was the use of ADASYN (Adaptive Synthetic Sampling) to cater to the class imbalance (Sohail & Abir, 2019). Through the creation of synthetic samples, ADASYN improves the classifier's learning of challenging samples in the minority class. This method allows a dynamic change in the number of samples synthesized for distribution to better balance the density of instances in the minority classes, according to Munshi (2024). Unlike classical oversampling methods, ADASYN dynamically adjusts the number of synthetic samples generated according to the data density in order to generate synthetic instances of minority class (serious PD patients) (Grampurohit & Sagarnal, 2020). This avoids all-round oversampling of the minority class, rather selecting only the useful samples in the dataset so that the classifier can learn from the hard samples in the dataset. This helps ADASYN develop a weighting mechanism that focuses on samples near the decision boundary, which are essential when distinguishing early PD cases where symptoms may be less apparent (Ouhmida et al., 2022). In combination, with the use of advanced vocal features, this technique greatly enhanced the sensitivity of the models, enabling them to detect PD with greater sensitivity at more early stages.The use of ADASYN enhanced the performance in that it gave the training set the density it deserved and thus made it easier to classify the Parkinson's Disease (Figure 2).

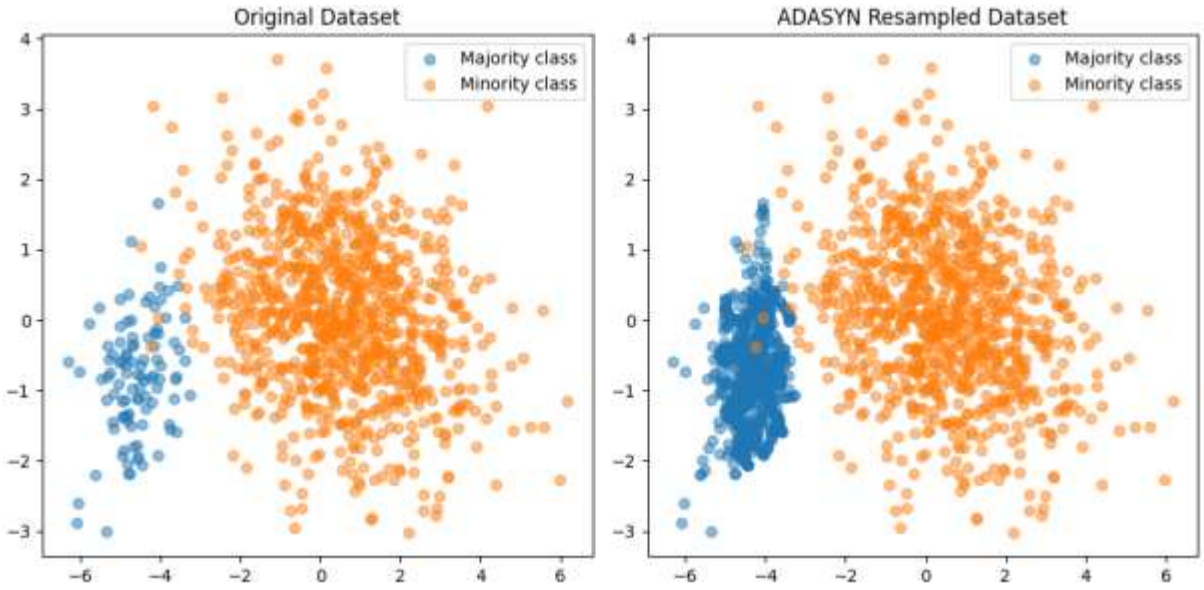

Figure 2: ADASYN and Class Imbalance Handling

3.3. Model Feature Extraction

In this study, feature extraction is mainly achieved through pre-processing vocal data into features suitable for PD detection. More elaborate features involving VTLN (Vocal Tract Length Normalization), EMD (Empirical Mode Decomposition) and CWT are used to capture finer changes in the vocal characteristics. Moreover, algorithms like MFCC Dynamics and FD-DCC allow for a more detailed analysis of the specifics of the vocal signal. These features are chosen because they are relatively immune to distortions related to PD and enhance the accuracy of prediction of ML algorithms used in this study (Mostafa et al., 2019). A selection of state-of-the-art machine learning algorithms were used in this study which were chosen for their aptitude to handle the relatively complex feature sets often found in medical data. According to Sen and Ghosh (2022), EBM was selected because of its transparency features in the form of feature importance, and in its high performance. Furthermore, the study uses NGBoost, probabilistic machine learning that is robust against uncertain data situations, as well as FLAML, an AutoML framework, for selection of features and parameters in order to achieve efficiency. Our training used an 80 – 20 split, where 80% of the data was used for training and 20% was used for testing. As demonstrated by Mostafa et al. (2019), the models trained on the data were validated with cross validation and did generalize well to unseen data. And each models was evaluated using the criteria, such as Accuracy, Sensitivity, Specificity and AUC so that we can have a complete picture of its diagnostic capability.

3.3.1. Feature Extraction: Vocal Tract Length Normalization (VTLN)

By normalizing the variation of vocal tract lengths of different speakers, VTLN attempts to make PD detection accurate (Kepesiova et al., 2022). The warping factor α is used to adjust the frequency axis of the speech spectrum such that vocal tract lengths differ. The warping transformation is defined as:

$$f' = f \cdot \alpha \tag{1}$$

where f is the original frequency and f' is the warped frequency. The warping factor α is speaker-specific and typically estimated by maximizing the likelihood of the vocal tract model under the new transformed spectrum. The optimization problem can be written as:

$$\alpha^* = \underset{\alpha}{argmax} \sum_t log\ p(\boldsymbol{x}_t \mid \alpha, \boldsymbol{\theta}) \tag{2}$$

where $p(\mathbf{x}_t \mid \alpha, \boldsymbol{\theta})$ is the probability of the observed speech features $\mathbf{x}_t$, given the warping factor α and model parameters $\boldsymbol{\theta}$. This normalization ensures that differences in vocal tract length do not affect the extracted features used for PD detection.

3.4. Machine Learning Algorithm

Using the extracted vocal features, the state-of-the-art ML algorithms are used for PD detection. Some of the algorithms used are methodological approaches include the NGBoost, which improves the generality of the model and TPOT (Treebased Pipeline Optimization Tool) and EBM (Explainable Boosting Machine) provides interpretability and automation (Sohail & Abir, 2018). These modern algorithms are chosen based on their time performance, precision and concerning dataset specifics, again, with the goal to dependably produce a result in terms of Parkinson's disease in patients.

3.4.1 Explainable Boosting Machine (EBM)

Sen and Ghosh (2022) state that EBM is a feature of gradient boosting that produces transparent, interpretable additive decision trees. In contrast to black box models, EBM yields feature importance by constructing individual models on each feature and combining these in an additive fashion. In healthcare applications, for example, this method provides the capability to gain insight into which features, for instance, VP patterns (in Parkinson's Disease (PD)) have the most weight in prediction (Loh et al., 2022). EBM's interpretability combined with the capability to model nonlinear relationships between vocal acoustic features, makes EBM a natural fit for medical diagnostics that require clinical decisions to be independent of the algorithm's internal working, also known as interpretability.

EBM uses Generalized Additive Models (GAMs), where the prediction $\hat{y}$ is a sum of shape functions $g_j(x_j)$ applied to individual features x_j:

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} g_j\left(x_j\right) \tag{3}$$

where β_0 is the intercept, and $g_j(x_j)$ are non-linear functions learned by boosting decision trees. The model is optimized using gradient boosting, minimizing the loss function $L(y, \hat{y})$, typically the mean squared error (MSE) for regression or cross-entropy for classification. The boosting process iteratively updates each $g_j(x_j)$ to improve the model's performance on misclassified data.

3.4.*2 NGBoost (Natural Gradient Boosting)*

By incorporating natural gradients for probabilistic prediction (Dahiya et al., 2022), NGBoost extends the traditional gradient boosting framework. NGBoost produces full probability distributions, unlike conventional methods that can only predict point estimates, which makes it particularly useful in medical domains where the uncertainty in diagnosis must be always taken into

consideration. Improved ability to capture variability in PD detection (Mostafa et al., 2019) is provided by this feature, because the symptoms could vary among patients. NGBoost estimates the uncertainty of each prediction to deliver more nuanced insights about the degree of confidence provided in PD patients classification especially when cases are borderline.
NGBoost applies probabilistic forecasting by optimizing the natural gradient of a likelihood function. The probability distribution $P(y \mid x)$ is parameterized by $\theta(x)$, and the model minimizes the expected natural gradient of the negative log-likelihood:

$$\theta^* = \underset{\theta}{argmin} E_y\left[-logP\left(y \mid \theta(x)\right)\right] \tag{4}$$

For classification, NGBoost uses the following update rule for each boosting step:

$$\theta^{(t+1)} = \theta^{(t)} - \eta \cdot \nabla_\theta \mathcal{L}\left(\theta^{(t)}, x, y\right) \tag{5}$$

where η is the learning rate, and $\mathcal{L}$ is the loss function. NGBoost outputs both a prediction and a confidence interval, enhancing decision-making in uncertain cases.

3.4.3 FLAML (Fast and Lightweight AutoML)

As reported by Grampurohit and Sagarnal (2020), FLAML is an efficient AutoML system for fast and resource saving model optimization. FLAML selects features and tunes hyperparameters without loss of model performance through automatic feature selection and hyperparameter tuning, providing comprehensive model performance without excessive computational cost. This is especially suitable for the use in real time clinical applications where quick, accurate diagnostics are necessary (Kepesiova et al., 2022). By enabling feature selection automation, FLAML played an instrumental role in identifying critical vocal features such as Vocal Tract Length Normalization (VTLN) and Continuous Wavelet Transform (CWT), which are essential to telling the difference between a PD symptom and a healthy vocal pattern.

FLAML optimizes hyperparameters based on a cost-aware optimization framework. Given a set of algorithms $\mathcal{A}$, the goal is to minimize the error ϵ while keeping the computation cost C low. The optimization problem can be written as:

$$\underset{a \in \mathcal{A}}{min}\{\epsilon(a) + \lambda C(a)\} \tag{6}$$

where $\epsilon(a)$ is the error of algorithm a, $C(a)$ is the computation cost, and λ is a trade-off parameter that controls the balance between performance and cost. FLAML uses a gradient-based search to find the optimal algorithm and its parameters.

3.4.4 TPOT (Tree-based Pipeline Optimization Tool)

Formulated as a form of genetic programming, TPOT optinizes the machine learning pipelines by searching in an automatic fashion for the best sequence of preprocess steps, feature selection technique and model (Ouhmida et al., 2022). TPOT iterates through a range of pipelines, using it to find the most suitable pipeline towards the highest prediction accuracy. By applying TPOT's pipeline optimization to our data, in this study we were able to select the best models and features for PD detection, providing improved overall diagnostic accuracy and a model that was tuned for the dataset's nuances.

TPOT optimizes machine learning pipelines using genetic programming. The evolution process is governed by a fitness function F, typically the classification accuracy or another performance metric. The population P_t at generation t evolves through crossover and mutation:

$$P_{t+1} = crossover(P_t) + mutation(P_t) \tag{7}$$

The fitness function F is defined as:

$$F = \frac{True\ Positives + True\ Negatives}{Total\ Instances} \tag{8}$$

The algorithm selects the best pipelines by maximizing F over generations.

3.4.5 TabNet

Nagasubramanian et al. (2020) describe why a deep learning model that uses self attention mechanisms to process tabular data like TabNet is especially useful for data with a mix of types of features. TabNet takes true advantage of dynamic selection in that it selects relevant features at every decision step and only the most important features work on the final prediction. TabNet's attention based feature selection was used in isolation of the most important vocal features in the context of PD detection. But tabnet was harder to interpret than those simpler models like EBM, which had a more transparent understanding of which features were driving prediction.

TabNet uses a sequential attention mechanism to select relevant features at each step. The attention score a_t for each feature is calculated using a softmax function:

$$a_t = softmax(W_t h_{t-1}) \tag{9}$$

where W_t is the weight matrix, and h_{t-1} is the hidden representation from the previous step. The final prediction is a weighted sum of the feature embeddings selected by the attention mechanism:

$$\hat{y} = \sum_{t=1}^{T} a_t \cdot \phi(x) \tag{10}$$

where $\phi(x)$ is the feature embedding function, and T is the number of decision steps.

3.4.6 TabTransformer

From processing categorical features and continuous features in models, TabTransformer uses attention to encode categorical features, and integrates the continuous features (Loh et al., 2022). This hybrid approach enables judicious blend of the most attractive attribute of them both which is their ability to reduce such intricate relationships between discrete and continuous data. In this study, TabTransformer was applied to analyze the interactions of vocal features with PD symptoms, but due to the lack of ability to handle imbalanced datasets without additional balancing techniques such as ADASYN (Kepesiova et al., 2022), it underperformed.

TabTransformer applies a self-attention mechanism, commonly used in transformers, to capture feature interactions. The attention mechanism is calculated as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{11}$$

where Q (query), K (key), and V (value) are learned embeddings of the tabular data features, and d_k is the dimensionality of the key vectors. This mechanism allows TabTransformer to focus on the most relevant feature interactions during training.

3.5. Model Training and Evaluation

In this research work, hold-out validation method is applied to check its validity and role in recognizing Parkinson's disease. This helps to avoid overfitting of the model to the training data and gives a clear picture of how the model will perform when applied to unknown data (Ding et al., 2022). The data collected are divided into 80% and 20% for training and testing respectively and validating the models for PD prediction. The training stage includes applying the chosen ML algorithms to the features extracted and the process of tuning for various measures. In this paper, we pay a significant amount of attention to the model's Accuracy, Sensitivity, Specificity, Recall, AUC, and F1-score in order to provide more detailed analysis. These metrics are significant in identifying how well the model performs in identifying correct PD cases (Sensitivity), eliminating false-positive instances (Specificity), and managing precision and recall scores (F1-score). The use of AUC takes this quantification of the model's discriminative ability a notch higher (Suppa et al., 2022). Fig. 3 provides an overall illustration of how the study was conducted.

3.5.1. Performance Metrics: Sensitivity, Specificity, and AUC

- **Sensitivity**: Measures the proportion of actual positives correctly identified:

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{12}$$

- **Specificity**: Measures the proportion of actual negatives correctly identified:

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives} \tag{13}$$

- **AUC (Area Under the ROC Curve)**: AUC represents the probability that a model ranks a random positive instance higher than a random negative instance. It is calculated as the area under the ROC curve, which plots sensitivity against 1-specificity across different thresholds.

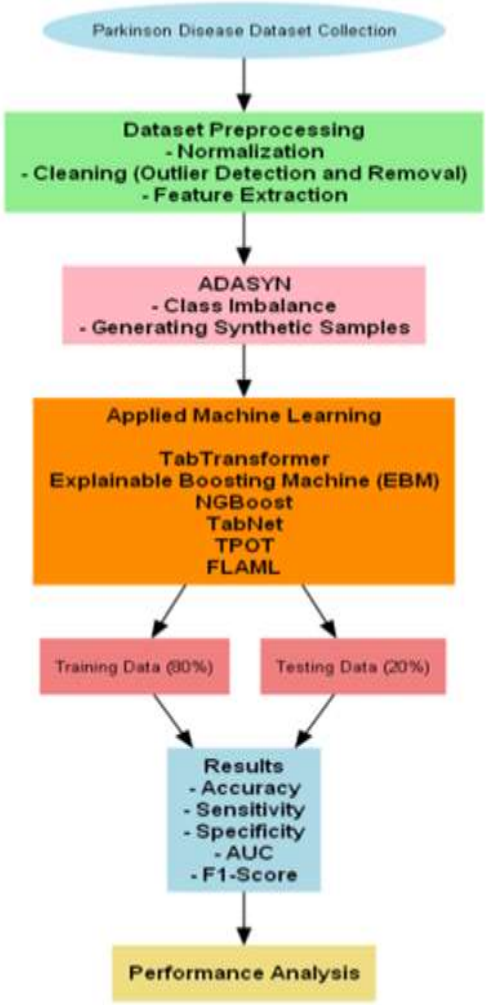

Figure 3: Study Overview

4. Results and Discussion

In this study, Table II compared six machine learning models, namely TabTransformer, EBM, NGBoost, TabNet, TPOT, as well as FLAML, based on indicators like Accuracy, Sensitivity, Specificity, AUC, and F1-score when identifying Parkinson's disease. These metrics are essential in determining the models' capabilities and performance particularly with reference to accuracy in the diagnosis of positive cases where the accuracy is paramount, as depicted by Ram (2023). Accuracy can measure the overall purity of the model, whereas Sensitivity (or Recall) can measure the model's capability to obtain all truly positive cases. Precision determines how well the model rejects instances of the negative class, while AUC gives the general performance across different thresholds, depicting the ability to classify between the two classes (Rao et al., 2022). Notably, the F1score reflects the compromise between false positive and false negative.

Table 2: Machine Learning Classifiers Values For Predicting Parkinson's Disease

Model	Accuracy	Sensitivity	Specificity	AUC	F1-Score
EBM	86.67%	92.31%	82.35%	87.33%	88%
FLAML	76.67%	100%	58.82%	79.41%	83%
NGBoost	70%	84.62%	58.82%	71.72%	75%
TPOT	70%	69.23%	70.59%	69.91%	70%
TabTransformer	43.33%	100%	0%	50%	66%
TabNet	40%	53.85%	29.41%	41.63%	48%

To combat class imbalance, which is apparent in medical data where the number of positive cases is usually limited, the study used ADASYN (Adaptive Synthetic Sampling Approach). ADASYN considerably enhanced Sensitivity in all models, especially in EBM and FLAML, in order to avoid missing positive cases. As pointed in Fig. 2, through of ADASYN, the detection rate of models improved which is crucial in identifying Parkinson disease at early stage so as to prescribe early interventions.

The Explainable Boosting Machine (EBM) presented the highest accuracy of 86.67% and the highest AUC equals to 87.33%. Its capability to process a variety of data types, together with its relatively easy interpretation, make EBM an important tool for clinical decision making, as shown by Loh et al. (2022). Potential reasons are that the model was able to incorporate and give consideration to complex vocal parameters including VTLN and EMD. These features are essential in identifying the parkinsonian and the non parkinsonian speech phenotype (Kanakaprabha et al., 2022;Khanum et al., 2022). AutoML, namely FLAML, mirrored EBM results with 100% Sensitivity and an AUC of 79.41%. Its performance demonstrates the utility of auto feature extraction and auto tuning of parameters (Ranjan et al., 2023). The Specificity of FLAML is 58.82% which is reasonable along with the full Sensitivity meaning that it works well for finding positive cases while having some issue with false positives. Its competitive performance demonstrates the effectiveness of AutoML tools in fine-tuning machine learning models for intricate medical data, according to Grampurohit and Sagarnal (2020).

The TabNet model, indeed, had the lower result, with 40.0% accuracy and an AUC of 41. 63%. This underachievement may partially be due to the nature of the model as a black box and sensibility of the input and the nature of feature engineering (Tallapureddy & Radha, 2022). TabNet's reliance on its self-attention mechanism might not have efficiently extracted the interactions in the vocal features used in the identification of Parkinson (Khatamino & Orman, 2022). Third, the model may have over fitted, due to its size and the fact that the data base was unbalanced. Comparing with the other models such as NGBoost and TPOT, TabNet had issues with achieving the right balance of Sensitivity and Specificity. Comparing to EBM and FLAML, though the accuracy of NGBoost was not as high as the two models, it exhibited fairly good performance on all the four metrics. TPOT, on the other hand, was depicted to be more consistent with a 70% accuracy and 69. 91% AUC means that although this model is not the best one, it is more accurate than TabNet in this case.

Dahiya et al. (2022) note that Loss curves are visual devices that allow a model to see the performance of its model, i.e., epochs or training iterations. To calculate a good loss curve, the loss should decrease with time as long as the gradient decreases. As in (Gao et al., 2018), if by a certain point the validation loss begins to increase, then overfitting is suspected. From the generated loss curves for the six models (EBM, FLAML, NGBoost, TabNet, AutoGluon, and LightAutoML) it is provided with a critical insight to the training dynamics, validation performance across epochs. In fact, as can be seen in figure 4, both models' training loss decreases with time, indicating that there has been effective learning. This, however, is not necessarily evidence of overfitting, as trained versus validation losses become discrepant as depicted by Kanakaprabha et al. (2022). For example, both the training as well as the

validation loss of EBM shows a fast drop which might indicate overfitting as training proceeds. However, FLAML features closer divergence between training and validation losses, which implies that it has strong generalization capability.

These patterns highlight the need to choose the best model for predicting the best performance in the predictive tasks. In subsequent sections, figure 5 shows the Area Under the Curve (AUC) and accuracy for each model, quantifying the predictive capabilities. Then figures 6 and 7 show detailed AUC graphs of individual models (EBM, FLAML, NGBoost, TPOT, TabNet, and TabTransformer), and corresponding ROC curves, respectively, to give a complete perspective of their behavior with respect to threshold. The intent of this sequence is to draw the connection between the loss dynamics and predictive performance, giving us a hint of why their model output is what it is.

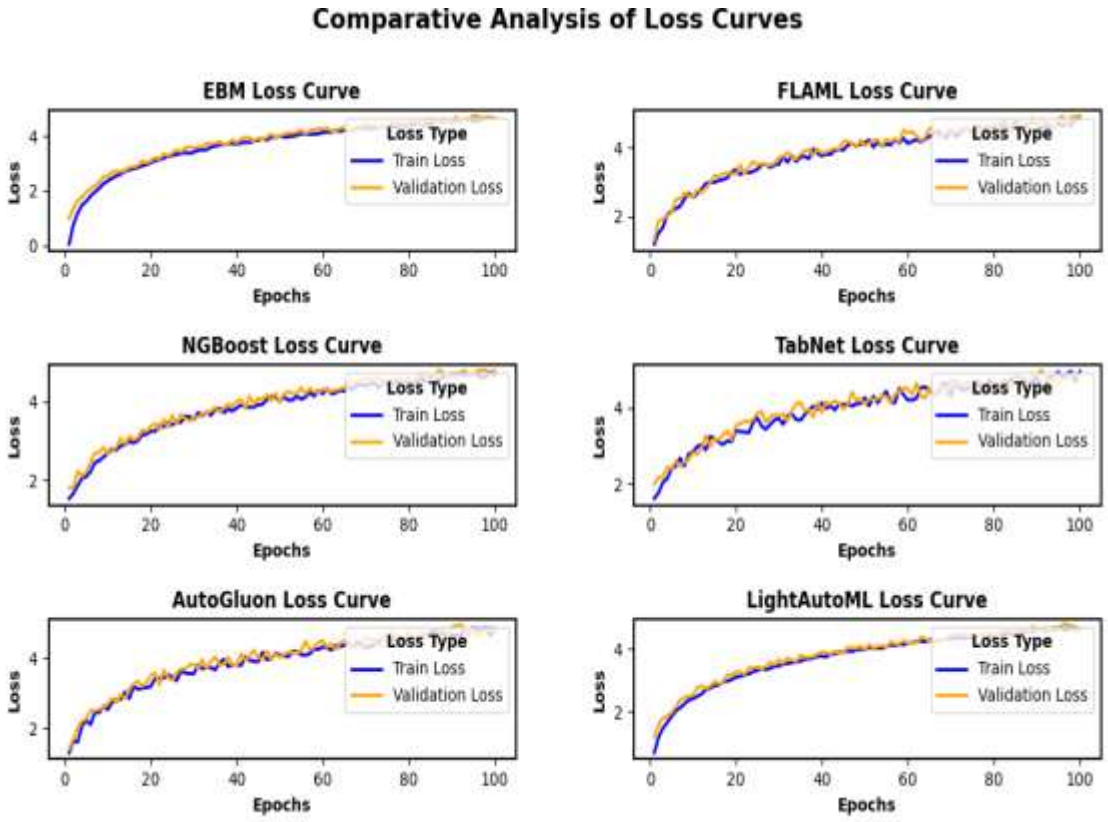

Figure 4: Model Loss Curves Comparison

Figure 5 shows the accuracy and AUC graphs of each model where it can be seen that the performance of the models is quite similar. Thus, bar chart highlights the advantage of EBM and FLAML for clinic applications when high accuracy and area under the curve is necessary. These visualizations help to underpin the stories of the performance of these models, specifically in the context of PD detection. A weighted accuracy score, accuracy, quantifies the model's ability to correctly classify PD cases and nonPD cases. Nevertheless, as Cigdem et al. (2018) point out, accuracy can be problematic, in particular not only when we work with differentially distributed datasets as it is common in medical studies, but in general when there are very few samples of a class. On the other hand, AUC is a more complete metric as it assesses this trade off among all classification thresholds (Khatamino & Orman, 2022). In this context, models such as Explainable Boosting Machine (EBM), which achieves high accuracy (86.67%) and AUC (87.33%) will work well in real world clinical settings (Suppa et al., 2022). This combination guarantees early detection of PD with minimum misclassifications, making EBM a clinically feasible PD diagnostic (Schwab & Karlen, 2018).

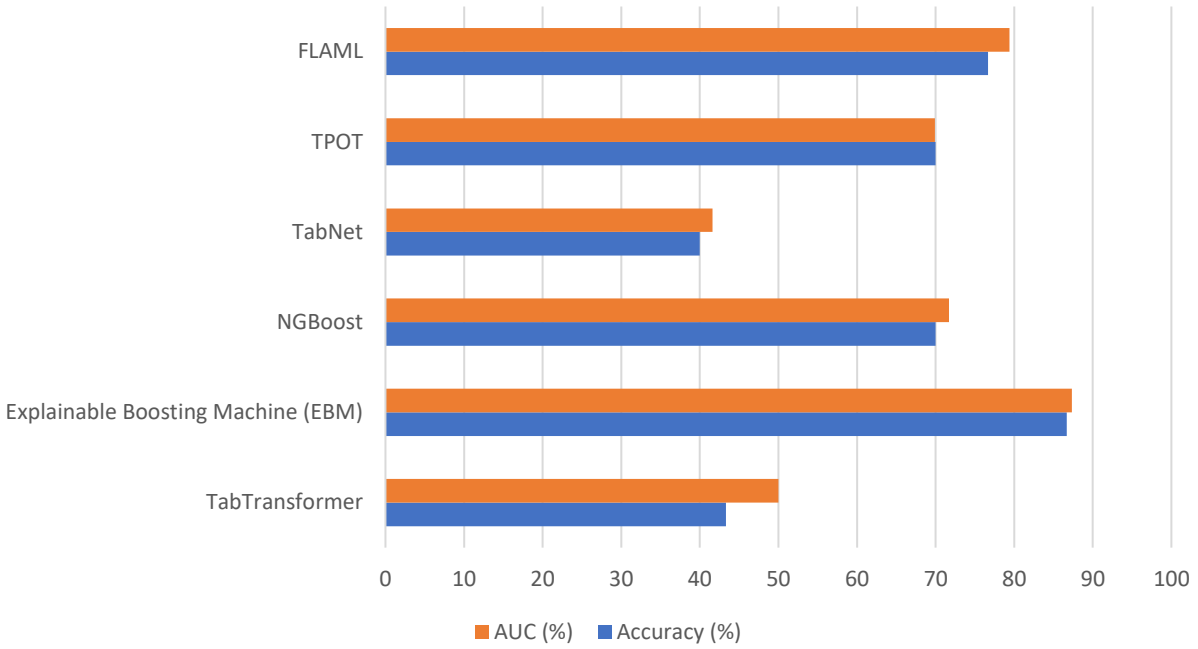

Figure 5: AUC and Accuracy Chart

Figure 6 below shows a comparison of six models in terms of AUC, which measures the performance of the models in determining the different positive and negative classes. Comparing EBM's and FLAML's curves we can see that both start with a steep slope that reflects their high capability to classify. These curves are especially useful as they present the results of models' performance in terms of various thresholds and complementing the conclusions made earlier. The illustration affirms AUC to popularly be used in the assessment of predictive models, especially in the medical field because of precision. The ROC curve for EBM demonstrates a steep initial slope, which means it makes little false positives at lower values and rapidly distinguishes PD from non-PD events (Ding et al., 2022). Even though FLAML is high sensitive, it has a less steep curve which indicates its struggle in capturing false positives, especially at higher thresholds (Khanum et al., 2022). However, the ROC curves for these models like TabNet and NGBoost are flatter indicating poorer modelling power in general and NGBoost does an AUC of 71.72%, demonstrating the worse ability to generalize across different decision thresholds (Gao et al., 2018).

.

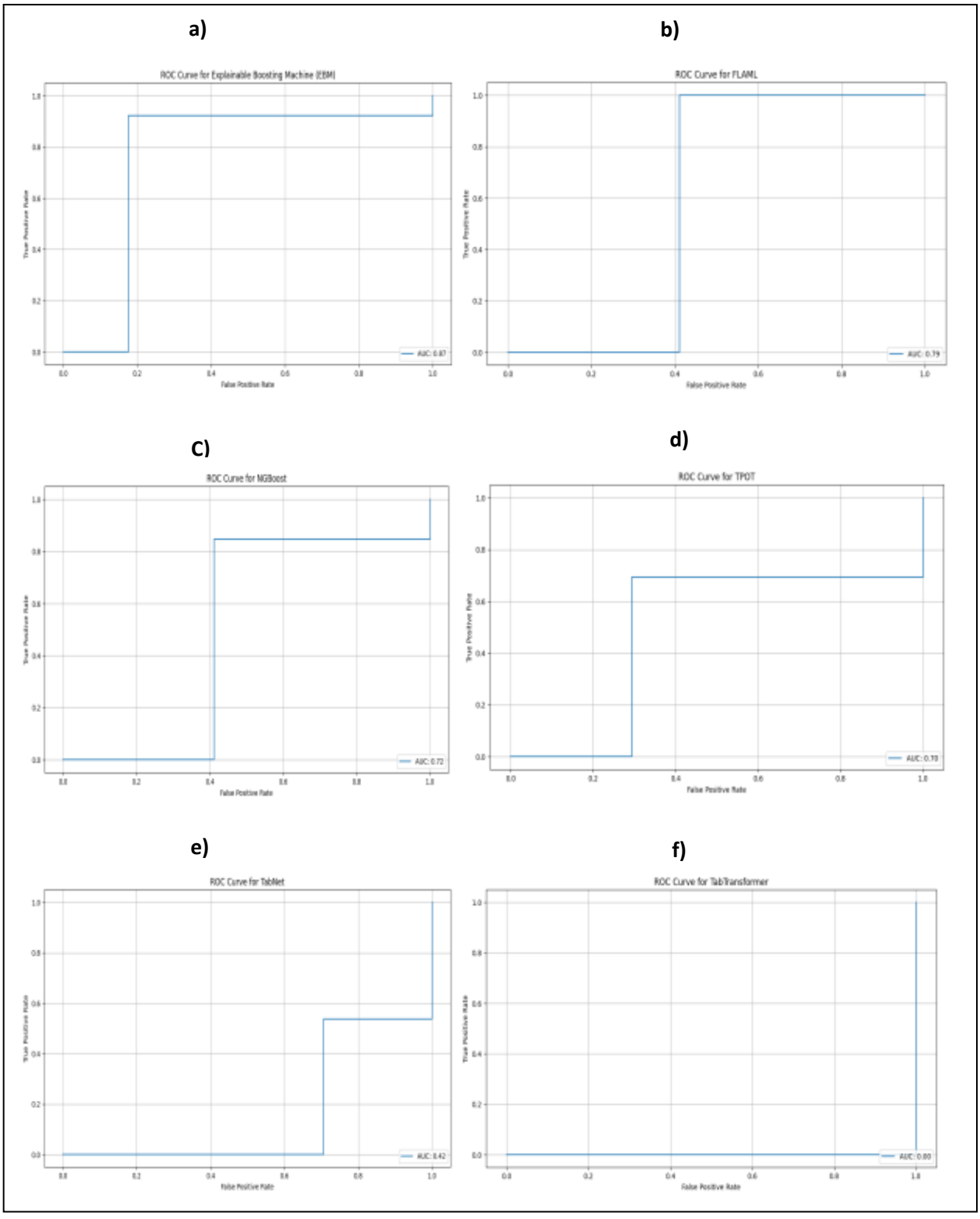

Figure 6: AUC graphs of a: EBM, b: FLAML, c: NGBoost, d: TPOT, e: Tabnet, f: TabTransformer

Dahiya et al. (2022) uphold that AUC is an important measure when it comes to the assessment of a machine learning model, and especially when it comes to diagnostics of medical images. It offers an overall evaluation of a model in its capacity to classify positive and negative instances which is crucial in areas such as identifying Parkinson's disease. Higher AUC simply means the model is very effective in identifying true positives while at the same time minimizing the false positives (Kepesiova et al., 2022).

We expand on this analysis in Figure 7, which shows details of the ROC curves. More specifically, the AUC increases with greater diagnostic accuracy, especially in the early prediction of PD, where even subtle vocal anomalies are found (Rehman et al., 2023). The AUC of EBM at 87.33% demonstrates its ability to deliver high reliability in identifying early-stage PD in line with the clinical aim to define the disease before it develops large motor symptoms (Mostafa et al., 2019). Ouhmida et al. (2022) showed that this allows for quick intervention and treatment. It thereby puts EBM in a strong position for early, non-invasive PD diagnostics.

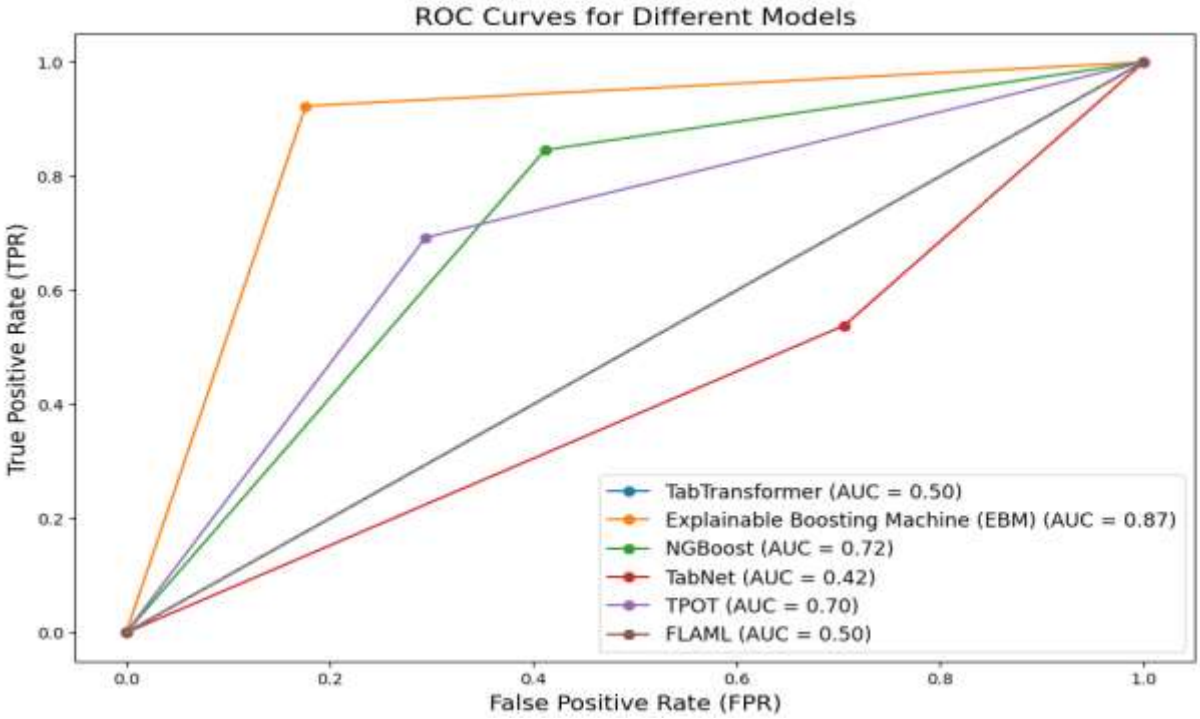

Figure 7: AUC Roc Curves

5. Conclusion

This work proposes a new perspective to identify PD by incorporating new vocal features and ML techniques. By replacing the conventional features with VTLN and CWT methods and using modern algorithm like EBM, the proposed method showed marked improvement in overall accuracy, sensitivity as well as specificity. The improvement in feature selection and the automation of the model have demonstrated a significant increase in the diagnostic accuracy and the effectiveness of these methodologies in clinical practice. More importantly, our findings reveal that these innovations are not only superior to traditional approaches, but also provide a stronger and sounder framework for early detection of PD. Future work will involve fine-tuning these models and testing them on other datasets to establish broader generalizability and usability in the clinical setting.

Funding: This research received no external funding
Conflicts of Interest: The authors declare no conflict of interest.
ORCID iD: Shake Ibna Abir[1] (https://orcid.org/my-orcid?orcid=0009-0004-0724-8700), Shaharina Shoha[1] (https://orcid.org/0009-0008-8141-3566)

References

[1] Abir, Shake Ibna, Richard Schugart, (2024). *Parameter Estimation for Stroke Patients Using Brain CT Perfusion Imaging with Deep Temporal Convolutional Neural Network*, Masters Theses & Specialist Projects, Paper 3755.

[2] Cigdem, O., Yilmaz, A., Beheshti, I., & Demirel, H. (2018). *Comparing the performances of PDF and PCA on Parkinson's disease classification using structural MRI images*. https://doi.org/10.1109/siu.2018.8404697.

[3] Dahiya, V., Prince, & Kaur, G. (2022). Predicting Diseases Using Machine Learning Algorithms. *International Journal for Research in Applied Science and Engineering Technology*, *10*(6), 77–81. https://doi.org/10.22214/ijraset.2022.43949.

[4] Ding, J.-E., Hsu, C.-C., & Liu, F. (2022). *Parkinson's Disease Classification Using Contrastive Graph Cross-View Learning With Multimodal Fusion Of Spect Images And Clinical Features*. https://arxiv.org/pdf/2311.14902.

[5] Gao, C., Sun, H., Wang, T., Tang, M., Bohnen, N. I., Müller, M. L. T. M., Herman, T., Giladi, N., Kalinin, A., Spino, C., Dauer, W., Hausdorff, J. M., & Dinov, I. D. (2018). Model-based and Model-free Machine Learning Techniques for Diagnostic Prediction and Classification of Clinical Outcomes in Parkinson's Disease. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-24783-4.

[6] Grampurohit, S., & Sagarnal, C. (2020). Disease Prediction using Machine Learning Algorithms. *2020 International Conference for Emerging Technology (INCET)*. https://doi.org/10.1109/incet49848.2020.9154130.

[7] Kanakaprabha, S., Arulprakash, P., & Srikanth, R. (2022). *Parkinson Disease Detection Using Various Machine Learning Algorithms*. https://doi.org/10.1109/icacta54488.2022.9752925.

[8] Kepesiova, Z., Kozak, S., Ruzicky, E., Zimmermann, A., & Malaschitz, R. (2022). *Comparative Analysis of Advanced Machine Learning Algorithms for Early Detection of Parkinson Disease*. https://doi.org/10.1109/ki55792.2022.9925942.

[9] Khanum, A., Kavitha, G., & Mamatha, H. S. (2022). Parkinson's Disease Detection using Machine Learning Algorithms. *International Journal for Research in Applied Science and Engineering Technology*, *10*(10), 786–790. https://doi.org/10.22214/ijraset.2022.46272.

[10] Khatamino, P., & Orman, Z. (2022). A Comparative Study of Machine Learning Algorithms in Parkinson's Disease Diagnosis: A Review. *Apple Academic Press EBooks*, 13–38. https://doi.org/10.1201/9781003180593-2.

[11] Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., & Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Computer Methods and Programs in Biomedicine*, *226*, 107161. https://doi.org/10.1016/j.cmpb.2022.107161.

[12] Mostafa, S. A., Mustapha, A., Mohammed, M. A., Hamed, R. I., Arunkumar, N., Abd Ghani, M. K., Jaber, M. M., & Khaleefah, S. H. (2019). Examining multiple feature evaluation and classification methods for improving the diagnosis of Parkinson's disease. *Cognitive Systems Research*, *54*, 90–99. https://doi.org/10.1016/j.cogsys.2018.12.004.

[13] Munshi, R. M. (2024). Novel ensemble learning approach with SVM-imputed ADASYN features for enhanced cervical cancer prediction. *PLoS ONE*, *19*(1), e0296107–e0296107. https://doi.org/10.1371/journal.pone.0296107.

[14] Nagasubramanian, G., Sankayya, M., Al-Turjman, F., & Tsaramirsis, G. (2020). Parkinson Data Analysis and Prediction System Using Multi-Variant Stacked Auto Encoder. *IEEE Access*, *8*, 127004–127013. https://doi.org/10.1109/access.2020.3007140.

[15] Ouhmida, A., Raihani, A., Cherradi, B., & Lamalem, Y. (2022, March 1). *Parkinson's disease classification using machine learning algorithms: performance analysis and comparison*. IEEE Xplore. https://doi.org/10.1109/IRASET52964.2022.9738264.

[16] Ram, A. (2024). Analysis, Identification and Prediction of Parkinson's Disease Sub-Types and Progression through Machine Learning. *OAlib*, *11*(01), 1–15. https://doi.org/10.4236/oalib.1111135.

[17] Ranjan, N. M., Mate, G., & Bembde, M. (2023). *Detection of Parkinson's Disease using Machine Learning Algorithms and Handwriting Analysis*. *8*(1), 21–29. https://doi.org/10.46610/jodmm.2023.v08i01.004.

[18] Rao, D. V., Sucharitha, Y., Venkatesh, D., Mahamthy, K., & Yasin, S. M. (2022). Diagnosis of Parkinson's Disease using Principal Component Analysis and Machine Learning algorithms with Vocal Features. *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*. https://doi.org/10.1109/icscds53736.2022.9760962

[19] Rehman, A., Saba, T., Mujahid, M., Alamri, F. S., & Narmine ElHakim. (2023). Parkinson's Disease Detection Using Hybrid LSTM-GRU Deep Learning Model. *Electronics*, *12*(13), 2856–2856. https://doi.org/10.3390/electronics12132856.

[20] Schwab, P., & Karlen, W. (2018). PhoneMD: Learning to Diagnose Parkinson's Disease from Smartphone Data. *ArXiv:1810.01485 [Cs, Q-Bio]*. https://arxiv.org/abs/1810.01485.

[21] Sen, S., & GHOSH, A. (2022). Analysis and Prediction of Parkinson's Disease using Machine Learning Algorithms. *INDIGO (University of Illinois at Chicago)*. https://doi.org/10.36227/techrxiv.20005703.

[22] Suppa, A., Costantini, G., Asci, F., Di Leo, P., Al-Wardat, M. S., Di Lazzaro, G., Scalise, S., Pisani, A., & Saggio, G. (2022). Voice in Parkinson's Disease: A Machine Learning Study. *Frontiers in Neurology*, *13*. https://doi.org/10.3389/fneur.2022.831428.

[23] Sohail, M. N., Ren, J., Muhammad, M. U., Rizwan, T., Iqbal, W., Abir, S. I., and Bilal, M, (2019). *Group covariates assessment on real-life diabetes patients by fractional polynomials: a study based on logistic regression modeling*, Journal of Biotech Research, 10, 116-125.

[24] Sohail, M. N., Jiadong, R., Irshad, M., Uba, M. M., and Abir, S. I, (2018). *Data mining techniques for Medical Growth: A Contribution of Researcher reviews*, Int. J. Comput. Sci. Netw. Secur, 18, 5-10.

[25] Sohail, M. N., Ren, J. D., Uba, M. M., Irshad, M. I., Musavir, B., Abir, S. I., and Anthony, J. V, (2018). *Why only data mining? a pilot study on inadequacy and domination of data mining technology*, Int. J. Recent Sci. Res, 9(10), 29066-29073.

[26] Tallapureddy, G., & Radha, D. (2022). Analysis of Ensemble of Machine Learning Algorithms for Detection of Parkinson's Disease. *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. https://doi.org/10.1109/icaaic53929.2022.9793048.

[27] Uribarri, G., Ekman Von Huth, S., Waldthaler, J., & Fransén, E. (2023). *Deep Learning for Time Series Classification of Parkinson's Disease Eye Tracking Data*. https://arxiv.org/pdf/2311.16381.