iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 Issue: III Month of publication: March 2024

DOI: https://doi.org/10.22214/ijraset.2024.59252

Comparative Analysis Using Machine Learning Algorithms to Detect Parkinson's Disease using Voice Dataset

Sk. Wasim Akram[1], Aamani Sandhya Tejaswini Mothadaka[2], Pervez Ahmed Shaik[3], Monshitha Manadala[4], Naga Siddhardha Pandi[5]
[1]Assistant Professor, [2, 3, 4, 5]UG Students, Department of CSE, Vasireddy Venkatadri Institute of Technology (Autonomous), Guntur, AP

***Abstract:** Parkinson's disease is a degenerative disorder that affects the control of movement, and its symptoms include tremors, stiffness, and difficulties in balance and coordination. This disease results from a decrease in dopamine-producing cells in the brain, and although there is currently no cure for Parkinson's, medication and therapeutic interventions can help alleviate the symptoms. Parkinson's disease usually begins to show its effects in middle to late adulthood. A new measure of dysphonia, Pitch Period Entropy (PPE) assesses the variability or randomness of pitch periods in speech signals, providing insights into the diversity of vocal characteristics. Higher entropy values suggest greater pitch variability in a speaker's voice. PPE is included in the dataset. The methodologies used in this project are Random Forest, Support vector classifier, Extreme gradient boosting (XGB). This comparative study includes performance metrics: Accuracy, precision, Recall, f1 score, and support. The aim is to predict Parkinson's disease and compare performance metrics using various methods.*

***Keywords:** Parkinson Disease, Dysphonia, tremors, Pitch Period Entropy, Entropy, Random Forest, Support Vector Machine,Extreme gradient boosting, Accuracy, precision, recall, f1-score.*

I. INTRODUCTION

Parkinson disease is a disorder related to neurons that leads to tremors, stiffness, and compromised equilibrium, exerting a substantial impact on an individual's motor faculties. This condition manifests through the gradual depletion of neurons responsible for dopamine production within the brain. Scientific inquiries have illuminated that upwards of 70% of individuals grappling with Parkinson's disease contend with voice and speech aberrations. Those afflicted with dysphonia amid this neurological encephalopathy frequently display vocal qualities marked by feebleness, harshness, or breathiness. Although Parkinson's disease can be detected by various symptoms such as tremors, stiffness, and impaired balance, an emerging focus in early diagnosis and monitoring revolves around analyzing voice patterns. Changes in voice characteristics, including pitch, tone,and rhythm, have been observed in individuals with Parkinson's. Leveraging voice-based technologies can provide a non- invasive and cost-effective method for detecting subtle changes in speech, potentially facilitating earlier diagnosis and more effective management of the disease.

II. LITERATURE REVIEW

There is less research on speech signals in Parkinson's disease because traditionally, the emphasis has been on motor symptoms, such as tremors and gait abnormalities. A previous study [2] used a Random Forest algorithm to perform early identification of Parkinson's disease based on speech signals. However, the acoustic features considered in the study were found to be unreliable in noisy environments. Another study [3] predicted PD with 85.06% accuracy using XG Boost algorithm. XG Boost is used to improve the accuracy in this project. There is a study titled "Speech Analysis for Diagnosis of Parkinson's Disease Using Genetic Algorithm and Support Vector Machine" [4].

However, the dataset used in this study contains unreliable features in noisy environments. In this project, Support Vector Machine (SVM) with a different dataset that has a better combination of features. Some studies have utilized handwriting analysis to identify Parkinson's disease by implementing deep learning techniques [5]. This paper proposes a computer vision-based system utilizing a fully convolutional neural network to recognize handwriting patterns. Their approach achieves a high accuracy of 92.43% in detecting Parkinson's patients, outperforming previous methods.

This research highlights the potential of computer-aided methods in diagnosing neurological disorders like Parkinson's Disease with improved precision and efficiency [6].

This paper reviews artificial intelligence techniques from 2016 to 2022 for Parkinson's disease screening and staging. It examines EEG, MRI, speech tests, handwriting exams, and sensory data for identifying PD biomarkers. The author discusses current and future trends in machine and deep learning for PD diagnosis, along with limitations and potential solutions. The aimis to enable early and precise diagnosis to initiate effective treatments and improve patients' quality of life [7]. This paper presents a novel approach to predict Parkinson's disease severity using deep neural networks on the UCI Parkinson's Telemonitoring Voice Data Set. Leveraging the TensorFlow library in Python, the methodology demonstrates improvedaccuracy compared to prior research efforts. By analyzing patient data, including speech patterns, the model aims to provide more accurate predictions of disease progression, aiding in better management and treatment planning for individuals with Parkinson's disease [8].

This study addresses the challenge of early Parkinson's disease detection using voice data collected from 31 patients. Imbalanced datasets were balanced using three sampling techniques. The proposed hybrid model achieved 100% accuracy, recall, and f1 score with random oversampling, and 100% precision, 97% recall, 99% AUC score, and 91% f1 score with SMOTE technique, showcasing promising results for improving diagnostic accuracy in Parkinson's disease detection [9]. This study proposes a deep learning approach to diagnose and predict the severity of Parkinson's disease using voice analysis, which is non-intrusive. By utilizing a mixed multi-layer perceptron (MLP) and autoencoder-based feature selection, it achieves a high accuracy of 99.15% in distinguishing severe from non-severe cases. Additionally, it accurately predicts the disease progression with a low Mean Squared Error (MSE) of 0.15, outperforming existing methods and offering promising prospects for early diagnosis and treatment monitoring of Parkinson ‘Detection of Parkinson’s Disease using Deep learning algorithms [10].

A combination of deep learning algorithms including Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN), along with a machine learning algorithm like Random Forest classifier, has been proposed to enhance the accuracy in identifying Parkinson's Disease (PD) from hand-drawn images. By leveraging these algorithms on a dataset comprising healthy and PD images, the accuracy rate has shown a significant improvement of 74%. This approach aids clinicians in more efficiently identifying and treating individuals with PD, potentially leading to better patient outcomes [11]. In this research, a novel approach combining Bag-of-Visual Words (BoVW) and Deep Optimum-Path Forest classifier is proposed for automatic Parkinson's disease identification. The method employs a hierarchical learning technique to construct visual dictionaries from handwriting exam data. Evaluation across six datasets demonstrates the effectiveness of the approach, highlighting its robust performance in computer-assisted medical diagnoses for Parkinson's disease detection [12]. Parkinson’s Disease (PD) affects millions globally, with early diagnosis crucial for improved patient outcomes. This research utilizes Deep Learning, specifically Convolutional Neural Networks (CNNs), to detect PD through analyzing Micrographia, a common symptom. By leveraging CNNs, the system achieves high accuracy (96.67%), precision, and recall, enhancing diagnostic accuracy and potentially improving the quality of life for PD patients. This approach showcases the potential of advanced technologies in early detection and management of neurological disorders like PD [13].

This study presents a Deep Convolutional Neural Network (DCNN) method utilizing MRI data for classifying Parkinson’s Disease (PD) stages, offering a promising approach for accurate diagnosis and staging of the disease. It explores variousnetwork structures to identify optimal accuracy and recall, trained on the Parkinson's Progression Markers Initiative (PPMI) database. Achieving 95% accuracy on PPMI data, the model outperforms traditional methods, offering potential for early-stage PD classification and progression prediction. This research highlights the promise of machine learning in enhancing PD diagnosis and management [14]. This article proposes a method for early prediction of Parkinson's disease (PD) using machine learning techniques with feature selection. It involves data preprocessing, feature extraction using MFEA, feature selection via PCA, and classification using a Darknet Convolutional Neural Network (DNetCNN). The proposed model achieves 97.5% accuracy in detecting PD early, outperforming existing methods and demonstrating superior performance in result evaluation[15].

III. PROPOSED METHODOLOGY

A. Dataset

Dataset consists of following 24 features:

TABLE 3.1Speech DataSet

name	object
MDVP:Fo(Hz)	Vocalic Pitch

MDVP:Fhi(Hz)	High-frequency Voicing
MDVP:Flo(Hz)	Low-frequency Voicing
MDVP: Jitter (%)	Microfluctuations
MDVP:Jitter(Abs)	Jitter Amplitude
MDVP:RAP	Rapid Jitter
MDVP:PPQ	Pitch Perturbation Quotient
Jitter:DDP	Jitter Dysregulation
	Vocalic Shimmer
MDVP:Shimmer(dB)	Shimmer Decibels
spread1	Dispersion1
spread 2	Dispersion2
MDVP:APQ	Vocalic Amplitude Perturbation Quotient
Shimmer:DDA	Shimmer Dysregulation Area
NHR	Noise-to-Harmonics Ratio
HNR	Harmonic-to-Noise Ratio
status	Int64 - 0/1
RPDE	Recurrence Period Density Entropy
DFA	Detrended Fluctuation Analysis
Shimmer:APQ3	Amplitude Perturbation Quotient 3
Shimmer:APQ5	Amplitude Perturbation Quotient 5
D2	Fractal Dimension
PPE	Pitch Period Entropy

B. Methods

In this project, XG Boost, Random Forest, and SVM are the primary methods.

1) XG Boost

XG Boost, an ensemble learning method, represents an intricately designed algorithmic paradigm renowned for its proclivity towards boosting predictive model accuracy.

Harnessing the amalgamation of decision trees, it iteratively refines the predictive capacity by emphasizing misclassified instances, thereby fortifying the model's discernment. This method, characterized by gradient boosting, exhibits a penchant for optimizing predictive performance, making it a prominent choice. The objective function in XG Boost is a combination of a loss function and a regularization term. It is optimized during the training process.

$$\text{Objective} = \sum_{i=1}^{n} \text{loss}(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

For binary classification, the logistic loss (log loss) is commonly used:

$$\text{Log Loss} = -\frac{1}{n} \sum_{i=1}^{n} [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)]$$

XGBoost includes a pruning strategy during the tree-building process to remove splits that do not contribute significantly to reducing the loss. This enhances the efficiency and prevents the model from becoming overly complex. XGBoost has a built-in mechanism to handle missing values in the dataset. It can automatically learn the best imputation strategy during training.

XGBoost is a versatile and powerful algorithm that has demonstrated state-of-the-art performance in many machine learning competitions and real-world applications. Its ability to handle complex relationships in data and its efficiency make it a go-to choose for many data scientists and machine learning practitioners. XGBoost excels in optimizing the trade-off between model complexity and generalization performance through the use of gradient boosting, regularization, and efficient tree construction. Its ability to handle diverse types of data, scalability, and interpretability make it a powerful algorithm in various machine tasks.

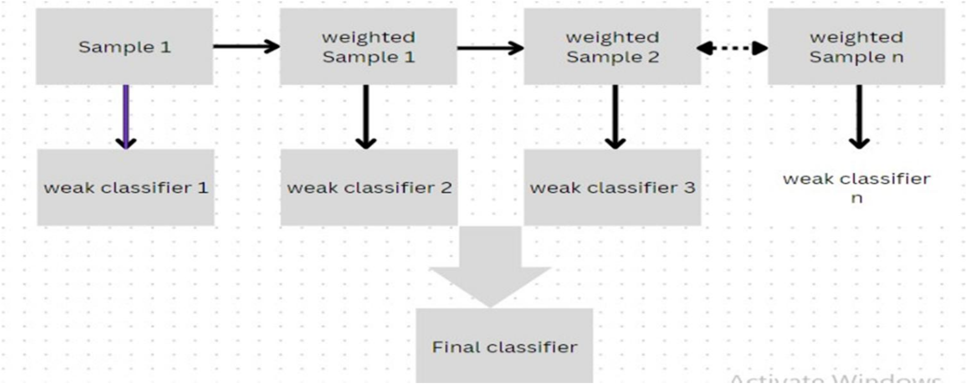

Fig 1 XG Boost working

2) Random Forest

Random Forest, a formidable ensemble learning paradigm, orchestrates a symphony of decision trees into a harmonious predictive melody. Distinctive in its arboreal diversity, it cultivates an assemblage of diverse trees by injecting randomness in feature selection and bootstrapping. This intricate orchestration curtails overfitting tendencies while enhancing predictive prowess, epitomizing a sylvan wisdom that navigates the intricacies of complex datasets with finesse and resilience. Random Forest operates through a collaborative and diverse assembly of decision trees. Random Forest starts by creating multiple subsets of the original dataset through bootstrap sampling. Each subset is a random sample with replacement. For each subset, a decision tree is constructed. However, at each node of the tree, only a random subset of features is considered for splitting. This injects diversity into individual trees. Each tree is grown by recursively partitioning the data based on the selected features. The split is determined by maximizing information gain or Gini impurity. When making predictions, each tree in the forest "votes" on the outcome. The predictions from all individual trees are aggregated to produce the final prediction. This ensemble averaging helps in reducing overfitting and improving generalization. The diversity among trees and the randomness injected during both sampling and feature selection contribute to the model's robustness. Random Forest tends to generalize well to new, unseen data.

The formula for the overall prediction in a Random Forest can be written as follows: If each tree in the forest outputs a class label, the final prediction is often the class with the majority of votes.

$$\hat{y}_{\text{RF}} = \text{mode}(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$$

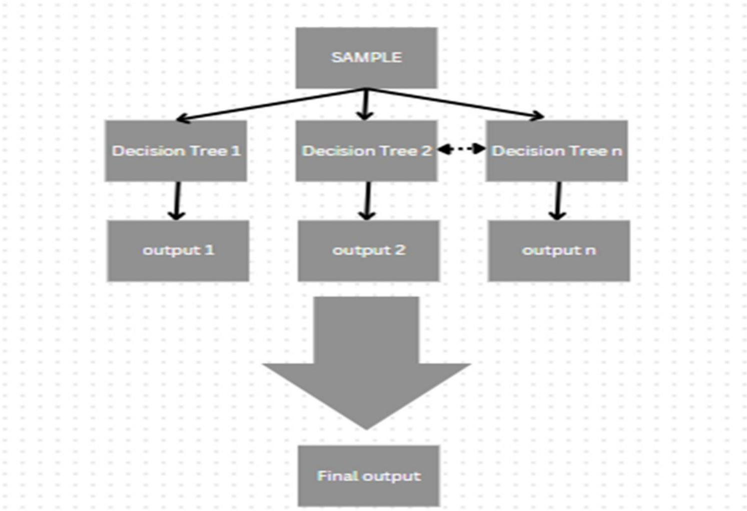

Fig -2: Random Forest working

3) Support Vector Machine

It works by finding a hyper plane in a high-dimensional space that best separates the data points into different classes. SVM can be likened to a discriminating function in a multidimensional space. Imagine a scenario where each data point is like a unique word, and the goal is to find the most effective way to draw a boundary between different groups of words. SVM identifies a hyper plane that maximizes the space between these word clusters, aiming to create a clear separation. This approach makes SVM robust in handling intricate patterns within data, even when dealing with limited instances of rare occurrences. The support vectors, representing pivotal words, play a crucial role in shaping the discriminating hyper plane, allowing SVM to make accurate predictions in complex, high-dimensional word spaces. The decision function, f(x), or a linear SVM is defined as the dot product of the feature vector x and the weight vector w, plus the bias term b:

Where , f(x) represents the decision function. w is the weight vector.
x is the input feature vector.
b is the bias term

The hyper plane equation is obtained by setting the decision function f(x) to 0.The margin represents the spatial separation between the hyper plane and the closest data point belonging to any class. The goal is to maximize this margin. The margin is given by the formula:

SVM involves optimization to find the optimal w and b. The objective function to be minimized subject to certain constraints is often expressed as:

Subject to the constraints :

$$y_i(f(x_i)+b) \geq 1 \quad \text{for all training samples } (x_i, y_i)$$

where y_i is the class label of the i-th sample.

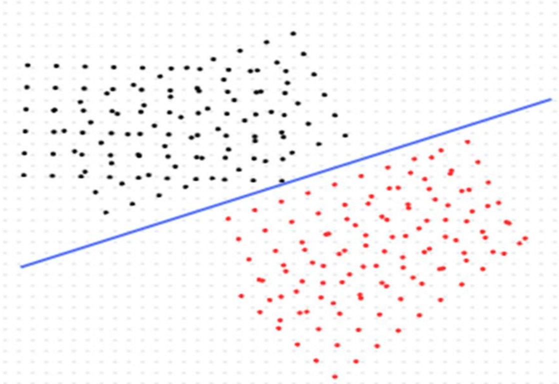

Fig. 4: Classification

IV. IMPLEMENTATION

A. *Preprocessing*

This project has a comprehensive dataset with no missing values. 'Standard Scalar' is used to scale the data.

B. *Model building and training*

1) *XG Boost classifier:* The training data is converted into DMatrix format. DMatrix is an internal data representation that optimizes the handling of sparse data, making it particularly effective for large datasets. The **objective** parameter in XG Boost is a crucial setting that defines the learning task and the corresponding objective function to be optimized during model training. It essentially specifies the type of predictive modeling problem you are working on. For binary classification problems, you typically set objective to "binary: logistic." This means XG Boost will optimize the logistic loss for binary classification tasks. The **eval_metric** parameter in XG Boost allows you to specify the evaluation metric to be used during the training of the model. This metric is different from the optimization objective defined by the objective parameter; it serves as a measure to assess the model's performance on the validation set during training. For binary classification, common choices include "error" (classification error rate), "logloss" (logarithmic loss), and "auc" (Area Under the ROC Curve). The eta parameter, also known as the learning rate, is a crucial hyper parameter in XG Boost that controls the step size or shrinkage during each iteration of the gradient boosting process. It is a value between 0 and 1 that determines the contribution of each tree to the final prediction. The value that has been taken is 0.1. A lower eta value makes the learning process more conservative, as each tree's contribution is scaled down. This helps prevent over fitting and can lead to a more robust model. The max_depth parameter in XG Boost determines the maximum depth allowed for each tree in the ensemble. In other words, it controls the depth of the individual decision trees within the boosting algorithm. The value taken in the project is 3. The subsample parameter in XG Boost controls the fraction of the training dataset that is randomly sampled and used to grow each tree during the boosting process. The subsample value in our model is 0.8. The colsample_bytree parameter in XG Boost controls the fraction of features (columns) randomly selected to grow each tree in the ensemble. The value taken for colsample_bytree is 0.8.

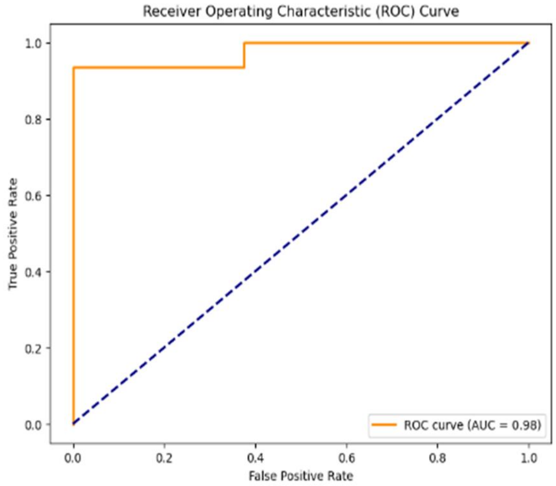

Fig 4.2.1 ROC Curve

2) *Random Forest classifier:* The data is transformed and used for training. One of the parameter of RandomForestClassifier is n_estimators the n_estimators parameter specifies the number of decision trees to be used in the random forest ensemble. It is a hyper parameter that you can tune during the model training process.
3) *SVM classifier:* This project has linear kernel to train the model in SVM, the parameter 'C' governs the balance between establishing a smooth decision boundary and accurately classifying the training points. It is essentially the inverse of regularization strength, where smaller values of 'C' result in a more regularized (smoother) decision boundary. The value of c=1.0.
4) *Model Evaluation:* The classification metrics that are used to evaluate models are:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

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

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

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

V. RESULTS & DISCUSSION

Algorithm Parameter	Xg Boost	Random Forest	svm
accuracy	0.948	0.923	0.871
precision	1.0	1.0	0.909
Recall	0.935	0.903	0.937
F1 score	0.966	0.949	0.923

Among the three models ,XG Boost has maximum accuracy of 94.87%.The next model that has maximum accuracy is RandomForest of 92.31%.The model that has least accuracy is SVM Model of 87.18%.Feature analysis is a crucial step in the machine learning pipeline that involves studying and understanding the importance and characteristics of the features (variables or attributes) used in a model.

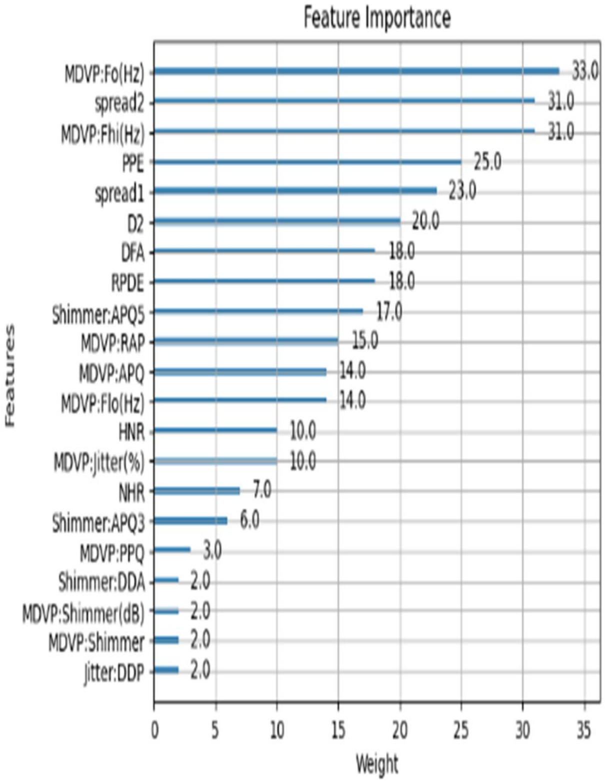

Fig 5.1 XG Model Feature Analysis

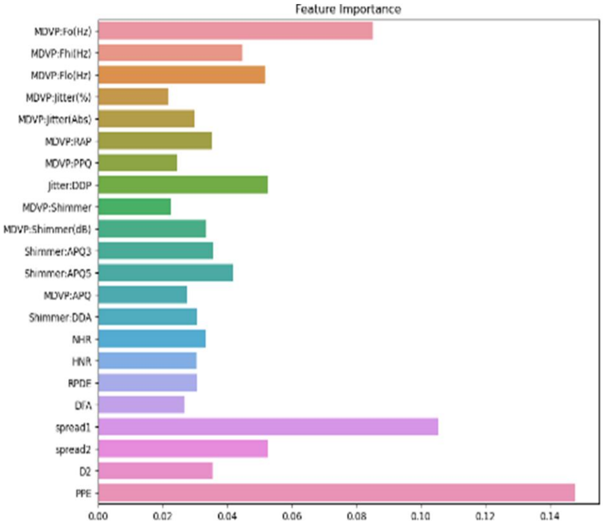

Fig 5.2 Random Forest Feature Analysis

Confusion Matrices by different models:

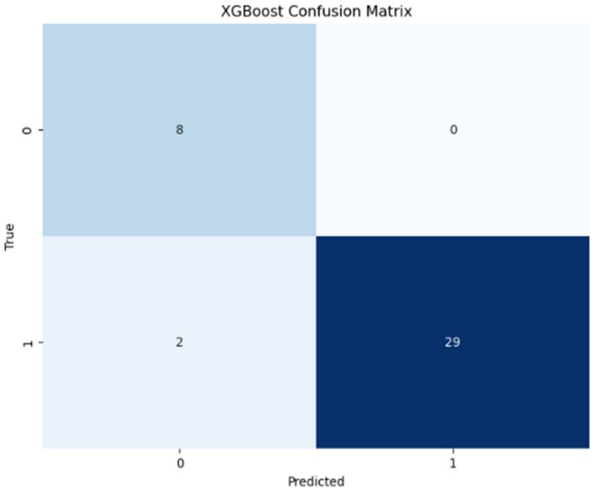

Fig5.3 XG Boost Confusion matrix

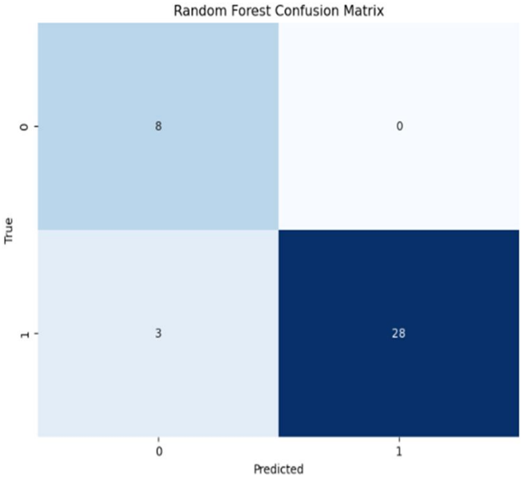

Fig 5.4 Random Forest Confusion matrix

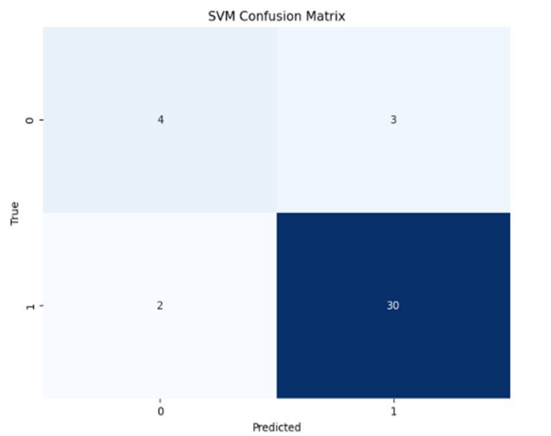

Fig 5.5 SVM Confusion matrix

VI. CONCLUSION

This investigation aimed to assess the performance of three well-known machine learning algorithms—XGBoost, Random Forest, and Support Vector Machines (SVM)—in a specific classification context. Through a rigorous examination and analysis process, findings consistently indicate that XGBoost exhibits superior predictive accuracy compared to both Random Forest and Support Vector Machines. The accuracy of XG Boost model is 94.87%.

VII. LIMITATIONS

This analysis has been conducted utilizing an easily accessible dataset. For a more comprehensive examination, researchers can consider collecting diverse data types by visiting hospitals. While the models employed in this study are robust, it is acknowledged that there is room for accuracy improvement. Notably, focus on a voice dataset for Parkinson's disease detection provides a foundational understanding. However, the integration of additional features such as smoking habits, educational background, and other behavioral attributes may contribute to more nuanced and refined results. Future studies could explore these unconsidered factors, potentially enhancing the overall accuracy and depth of Parkinson's disease detection models.

VIII. FUTURE WORK

Future analyses could extend beyond the current dataset by incorporating a broader range of socio-demographic features. Enhancing accuracy remains a priority, and efforts can be directed towards refining model training methodologies. The exploration of an expanded repertoire of models could further enrich analytical insights, fostering a more comprehensive understanding of the subject matter. These considerations signify potential avenues for improvement in subsequent research endeavors.

IX. ACKNOWLEDGMENT

We hereby acknowledge that we have undertaken this project leveraging our knowledge and skills. Throughout the course of our work, we have responsibly consulted various reputable sources available on the internet to enrich our understanding and ensure the thoroughness of our project. The incorporation of insights from these references has contributed to the depth and quality of our research. Gratitude is extended to our project guide and counselor for their invaluable guidance and support throughout the duration of this project. Their expertise and encouragement have played a pivotal role in shaping our research journey and fostering our intellectual growth.

REFERENCES

[1] Ali L, Javeed A, Noor A, Rauf HT, Kadry S, Gandomi AH. Parkinson's disease detection based on features refinement through L1 regularized SVM and deep neural network. Sci Rep. 2024 Jan 16;14(1):1333. doi: 10.1038/s41598-024-51600-y. PMID: 38228772; PMCID: PMC10791701.

[2] María Teresa García-Ordás, José Alberto Benítez-Andrades, Jose Aveleira-Mata, José-Manuel Alija-Pérez & Carmen Benavides. “Determining the severity of Parkinson’s disease in patients using a multi-task neural network”, June 2023

[3] Multimedia Tools and Applications 83(2) DOI:10.1007/s11042-023-14932-x

[4] Little MA, McSharry PE, Hunter EJ, Spielman J, Ramig LO. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Trans Biomed Eng. 2009 Apr;56(4):1015. doi: 10.1109/TBME.2008.2005954. PMID: 21399744; PMCID: PMC3051371.

[5] Ping Fan, "[Retracted] Random Forest Algorithm Based on Speech for Early Identification of Parkinson'sDisease", Computational Intelligence and Neuroscience, vol. 2022, Article ID 3287068, 6 pages, 2022.

[6] G Abdurrahman, M Sintawati, Taheri Far1, Ehsan Tahami2, "Speech Analysis for Diagnosis of Parkinson'sDisease Implementation of xgboost for classification of parkinson's disease", i 2020 J. Phys.: Conf. Ser. 1538 012024.

[7] Mohammad Shahbakhi, Danial Taheri Far, Ehsan Tahami Department of Biomedical Engineering, Dezful Branch, Islamic Azad University, Dezful, Iran.Department of Biomedical Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran. "Speech Analysis for Diagnosis of Parkinson's Disease Using Genetic Algorithm and Support Vector Machine". Journal of Biomedical Science and Engineering,March 2014

[8] Minhazul Arefin , Kazi Mojammel Hossen, Rakib Hossen and MohammedNasirUddin. "Parkinson's Disease Handwriting Detection using Fully Convolutional Neural Network", EasyChair Preprint,March,2022.

[9] Mohamed Shaban,"Deep Learning for Parkinson's Disease Diagnosis",Computers 2023, 12(3), 58; https://doi.org/10.3390/computers12030058,February,2023.

[10] Srishti Grover, Saloni Bhartia, Akshama, Abhilasha Yadav, Seeja K.R."Predicting Severity Of Parkinson's Disease Using Deep Learning", Procedia Computer Science,Volume 132, 2018, Pages 1788-1794,2018.

[11] Amjad Rehman 1ORCID,Tanzila Saba 1ORCID,Muhammad Mujahid 2,Faten S. Alamri 3,*ORCID and NarmineElHakim,"Parkinson's Disease Detection Using Hybrid LSTM-GRU Deep Learning Model", Electronics 2023, 12(13), 2856; https://doi.org/10.3390/electronics12132856, June,2023.

[12] THANDAVAMEGANATHAN*, KRISHNAN," Micrographia-based parkinson's disease detection using DeepLearning", EasyChair Preprint, 2023.

[13] Dr.A.Christy Jeba Malar, Shivani BalajiSrivastava2,SriRavi,TinkuRam" Detection of Parkinson's Disease using Deep learning algorithms", E3S Web Conf.Volume 491, 2024

[14] Raziya Begum, Thummala Pavan Kumar, Manda Rama Narasinga Rao," Deep Convolutional Neural Networks for Diagnosis of Parkinson's Disease Using MRI Data", IIETA,2023.

[15] Luis Claudio Sugi Afonso,Clayton Pereira,Silke Anna Theresa Weber,Christian Hook," Hierarchical Learning Using Deep Optimum-Path Forest", Journal of Visual Communication and Image Representation,May,2020.

[16] G. Prema Arokia Mary1,* and N. Suganthi," Detection of Parkinson's Disease with Multiple Feature Extraction Models and Darknet CNNClassification",Tech Science Press, DOI:10.32604/csse.2022.021164,2021