Advance Journal of Engineering Research (IAJER)
*Volume 08, Issue 05 (**May - 2025**), PP 09-16*
ISSN: 2360-819X
www.iajer.com

Research Paper **Open Access**

An Integrative Approach to Diagnosing Parkinson's Disease using Ensembled ML Techniques

G.Harichandana [1], Y.V. Subramanyam[2], V. Kesava Kumar [3]
[1](Student, Dept. of. Computer Science and Engineering, Sri Mittapalli College of Engineering,Guntur, India)
[2](Assoc.Prof, Dept. of. Computer Science and Engineering, Sri Mittapalli College of Engineering,Guntur, India)
[3](Assoc.Prof, Dept. of. Computer Science and Engineering, Sri Mittapalli College of Engineering,Guntur, India)
**Corresponding author: G. Harichandana*

ABSTRACT : *A neurodegenerative disease that affects the brain's neurological, physiological, and behavioral systems, Parkinson's disease (PD) is difficult to diagnose early because of its subtle symptoms. Slowness of movement, or bradykinesia, is a hallmark PD symptom that usually first appears in middle adulthood and gradually worsens. Verbal communication impairment is one of the major effects of Parkinson's disease. Support Vector Machine (SVM), Naïve Bayes, k-Nearest Neighbors (k-NN), and Artificial Neural Networks (ANN) were among the supervised classification techniques used in this study to identify speech-related abnormalities linked to Parkinson's disease. Using the combined predictive strength of several models that were trained separately, ensemble learning techniques were used to improve diagnostic robustness. In particular, the following four ensemble algorithms were compared: majority voting, weighted voting, bagging, and AdaBoost. The 195 speech signal samples in the dataset included 48 samples from healthy controls and 147 samples from people with Parkinson's disease. The results of the experiments showed that ensemble approaches performed noticeably better than individual classifiers. Using a majority voting ensemble that integrated Decision Tree, k-NN, and SVM classifiers, the highest accuracy of 95.3 percent was attained, highlighting the effectiveness of ensemble techniques in improving PD detection of speech patterns.*

Keywords *- neurodegenerative disorder, Parkinson's disease; machine learning, disease prediction, adaboost, bagging, majority voting, soft voting, ensembled.*

I. INTRODUCTION

The Parkinson's disease (PD) is a progressive neurological disorder characterized by motor symptoms such as tremors, rigidity, bradykinesia, and postural instability. One of the fundamental pathological features of PD is the depletion of dopamine levels in the brain, which adversely affects both motor and physical functioning. Globally, PD is recognized as one of the most prevalent neurodegenerative diseases. The neurological symptoms associated with PD emerge sporadically and tend to intensify over time due to the progressive nature of neuronal damage [1].

Aging significantly contributes to the onset and progression of PD, primarily due to structural and functional changes in the brain, including the reduction of synaptic connections and alterations in neurotransmitter and neurohormone levels. As individuals age, neuronal regeneration diminishes, exacerbating the deterioration of brain function. These neurological and biochemical changes often remain undetected until the disease has advanced to a stage requiring medical intervention [2].

The manifestation of PD symptoms varies significantly among individuals. Common signs include impaired speech, memory loss, imbalance, and abnormal posture [3]. According to a 2019 report by the World Health Organization (WHO), approximately 8.5 million individuals are diagnosed with PD annually, making it the second most common neurodegenerative disorder after Alzheimer's disease [4, 5]. Although PD predominantly affects the elderly, around 4% of diagnosed individuals are under the age of 50 [4].

At present, no cure for PD exists. Clinical management is primarily symptomatic, with no therapeutic interventions available to halt or reverse disease progression [6]. Additionally, the absence of a definitive diagnostic test for PD means that diagnosis heavily relies on clinical history and the presentation of symptoms [7]. Given that existing diagnostic procedures are often invasive, costly, and logistically demanding, the development of accessible and reliable diagnostic alternatives is of paramount importance [8].

In recent decades, machine learning (ML)—a subfield of artificial intelligence (AI)—has emerged as a promising tool for the early detection and diagnosis of PD. ML models, when used alongside conventional

diagnostic practices, have demonstrated potential in enhancing diagnostic accuracy [9]. Among the physiological indicators of neurological disorders, gait abnormalities have been identified as significant due to their frequent manifestation in daily activities. Gait analysis, being non-invasive, presents an attractive diagnostic alternative, particularly in home settings [10, 11]. Some studies have explored the offline implementation of multiple ML algorithms to automate the diagnostic process [12, 13].

Speech impairments are also frequently observed in the early stages of PD. Disorders such as hypophonia (reduced vocal loudness), dysphonia (disrupted vocal quality), and echolalia (involuntary repetition of words) are notable vocal anomalies associated with the disease. Modern computational tools can analyze acoustic data derived from human speech, offering another avenue for non-invasive PD diagnosis [14].

II. LITERATURE SURVEY

This research distinguishes itself from previous studies by developing diagnostic systems capable of accurately differentiating between individuals with Parkinson's disease (PD) and healthy subjects through the analysis of speech data using a wide array of machine learning (ML) tools and approaches. To enhance diagnostic accuracy, M. AI-Sarem et al. [15] proposed a multi-algorithmic framework involving CatBoost, Random Forest (RF), and Extreme Gradient Boosting (XGBoost) for the diagnosis of PD. Their comparative analysis of these ensemble classifiers provided insights into the performance differentials of these advanced ML models.

A notable contribution to audio-based diagnosis was made by T. J. Wronge et al. [16], who introduced a Voice Activity Detection (VAD) approach for predicting PD. In their methodology, raw audio recordings underwent preprocessing steps to eliminate background noise. Subsequently, two distinct feature extraction algorithms were applied, followed by classification using ML techniques.

Additionally, K. R. Wan et al. [17] focused on function selection (FS) in the context of neurosurgery, specifically during surgical procedures involving PD patients. Their study applied ML algorithms to accurately localize the anatomical regions targeted for intervention, thus emphasizing the post-diagnosis application of ML in PD treatment.

Motion-based detection of PD was explored by Cavallo et al. [18], who collected kinematic data from the upper limbs of both PD patients and control subjects. Using implanted motion sensors, participants were instructed to perform a range of physical tasks. The collected data were analyzed via spatiotemporal and frequency domain techniques, followed by classification using multiple supervised learning algorithms.

Further, J. S. Almeida et al. [19] employed various feature extraction and ML methodologies to identify PD, particularly emphasizing phonation as a principal feature for diagnosis. Their comparative study evaluated the performance of classifiers such as k-Nearest Neighbors (K-NN), Multilayer Perceptron (MLP), Optimum Path Forest, and Support Vector Machine (SVM). Similarly, Parisi et al. [20] enhanced speech-based PD identification by applying Artificial Neural Networks (ANNs) for dimensionality reduction, followed by SVM for classification.

Beyond speech and motion data, researchers have also investigated handwriting analysis as a diagnostic modality for PD, exploring its utility as a behavioral biomarker [21].

While these prior studies have reported promising classification accuracy, they often suffer from drawbacks such as the use of extensive feature sets—resulting in increased computational costs—or difficulty in extracting reliable features even from minimal datasets. These challenges underline the need for optimized feature extraction techniques that balance efficiency and accuracy.

Accordingly, the current study aims to reduce computational complexity by identifying a minimal yet highly informative set of speech features for classification. Compared to methods based on magnetic resonance imaging (MRI) or motion data, speech-based feature acquisition is both non-invasive and easier to implement, making it highly suitable for scalable and accessible diagnostic systems. The central goal of this research is to explore machine learning approaches for the early prediction of Parkinson's disease using optimized speech-derived features.

III. PROPOSED METHODOLOGIES

In this section, we will cover the features extraction approaches, the traditional classifiers that are used for assembling, which are also known as base classifiers, as well as the ensemble classifier that has been recommended. By utilizing the Speed up Robust Features techniques, the initial part of the procedure involves the extraction of the features that are contained inside the dataset. The values of the extracted features are then entered into the basic classifiers, which can include types such as SVM, DT, LR, RF, MLP, NB, and k-NN. After that, the extracted features are used to classify the data. A better degree of accuracy in the diagnosis of PD disease is achieved by the utilization of the ensemble technique, which involves the combination of the three most effective base classifiers. As shown in Figure 1, the proposed PD diagnosis approach is consisting of following stages:

- Dataset Collection stage
- Pre-processing stage
- SIFT Feature extraction stage
- Traditional Classification
- Integrative Ensembled Techniques.

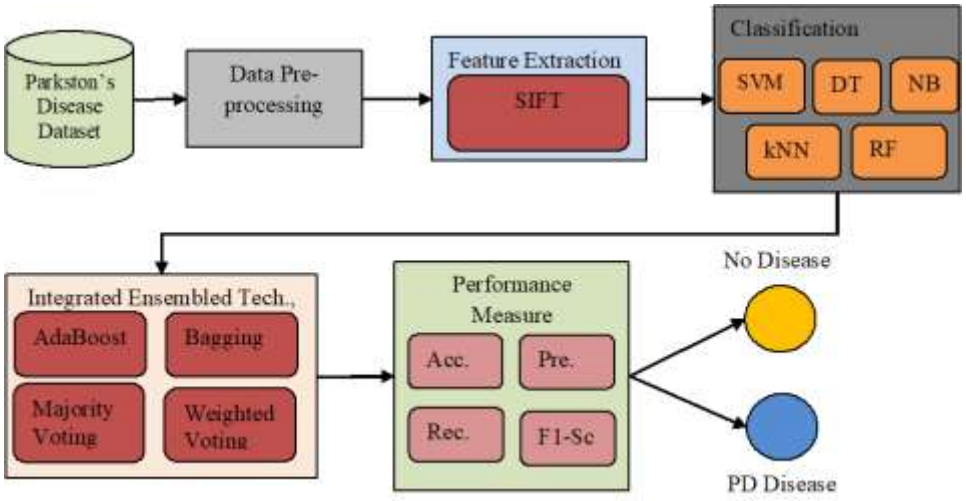

Fig. 1 Architecture of Parkinson disease diagnosis

A. Dataset Collection

The dataset used for early PD detection in this article is based on speech sounds, which is known as Parkinson's disease (PD). Max Little of Oxford University created it and donated it to the UCI Machine Learning Repository. The data set is widely considered by medical experts to be among the most efficient ever collected, organized, and analyzed. Many researchers have developed and evaluated automated algorithms using this dataset. Many researchers and others who are interested in early PD diagnosis still hold this objective in high regard. There are 195 biological sounds in the speech signal collection; 48 of them are considered healthy, while 147 are exclusive to patients with Parkinson's disease (PD) [22]. The voice measure and its interpretation are described in the first table, which presents 23 features derived from speech signals. Figure 2 highlights the aspects of the patient information related to Parkinson's disease.

The information contained in the Parkinson disease dataset is divided into two distinct portions, namely for training and testing purposes. It is demonstrated in Figure 3 that 75% of the datasets pertaining to PD diagnosis are utilized for training purposes, while the remaining 25% are utilized for testing purposes.

B. Dataset Processing

The processing of vast volumes of data is necessary to make sense of the information. The use of data analytics is one of the most important steps that must be taken to ensure that future initiatives will be effective.

#	Column	Non-Null Count	Dtype
0	name	195 non-null	object
1	MDVP:Fo(Hz)	195 non-null	float64
2	MDVP:Fhi(Hz)	195 non-null	float64
3	MDVP:Flo(Hz)	195 non-null	float64
4	MDVP:Jitter(%)	195 non-null	float64
5	MDVP:Jitter(Abs)	195 non-null	float64
6	MDVP:RAP	195 non-null	float64
7	MDVP:PPQ	195 non-null	float64
8	Jitter:DDP	195 non-null	float64
9	MDVP:Shimmer	195 non-null	float64
10	MDVP:Shimmer(dB)	195 non-null	float64
11	Shimmer:APQ3	195 non-null	float64
12	Shimmer:APQ5	195 non-null	float64
13	MDVP:APQ	195 non-null	float64
14	Shimmer:DDA	195 non-null	float64
15	NHR	195 non-null	float64
16	HNR	195 non-null	float64
17	status	195 non-null	int64
18	RPDE	195 non-null	float64
19	DFA	195 non-null	float64
20	spread1	195 non-null	float64
21	spread2	195 non-null	float64
22	D2	195 non-null	float64
23	PPE	195 non-null	float64

Fig. 2 Attribute information of PD disease

To begin, it is necessary to produce data by completing any missing values, removing any outliers, and removing any duplicates. Second, it is essential to validate the data to ensure its completeness and consistency. Since there are as many rows as there are unique column values, we found that the dataset does not include duplicate values in this article. All of the features are continuous "numerical variables" kinds, with the exception of the "status" feature, which is a binary categorical kind. The kind and degree of the faults detected in the data processing operations dictate the actions to be taken to rectify them. Some examples of what this involves are filling in missing numbers, handling outliers, removing duplicates, and investigating and correcting data-validation mistakes. For reliable research, it is essential to ensure that the data is both full and of high quality. Figure 4 shows the Parkinson disease dataset with all of its variables and the range of possible values for each.

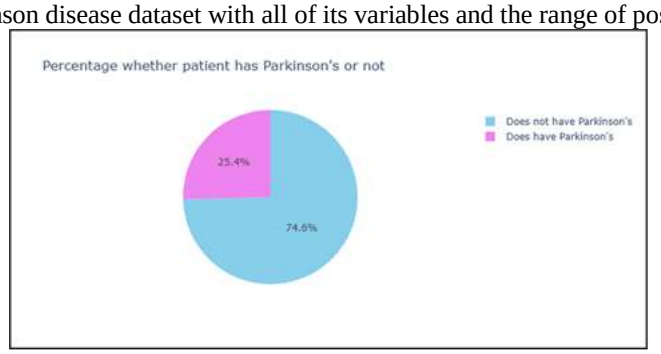

Fig. 3 PD disease dataset split up

C. SIFT Feature Extraction

David Lowe was the first to introduce the scale invariant feature transform [23]. This is accomplished by means of an image search for interesting locations and the provision of local descriptions that shed light on the immediate area. After processing the image using the Difference of Gaussian (DoG), the first step of this method is to find extrema in the image. It filters the input picture at different sizes and then gradually decreases the sample size to achieve scale invariance. After that, we compare the pixels. It is also examined along with the lower and higher levels that are immediately next to it. Possible key points are adjacent pixels that are either the highest or lowest value in the set.

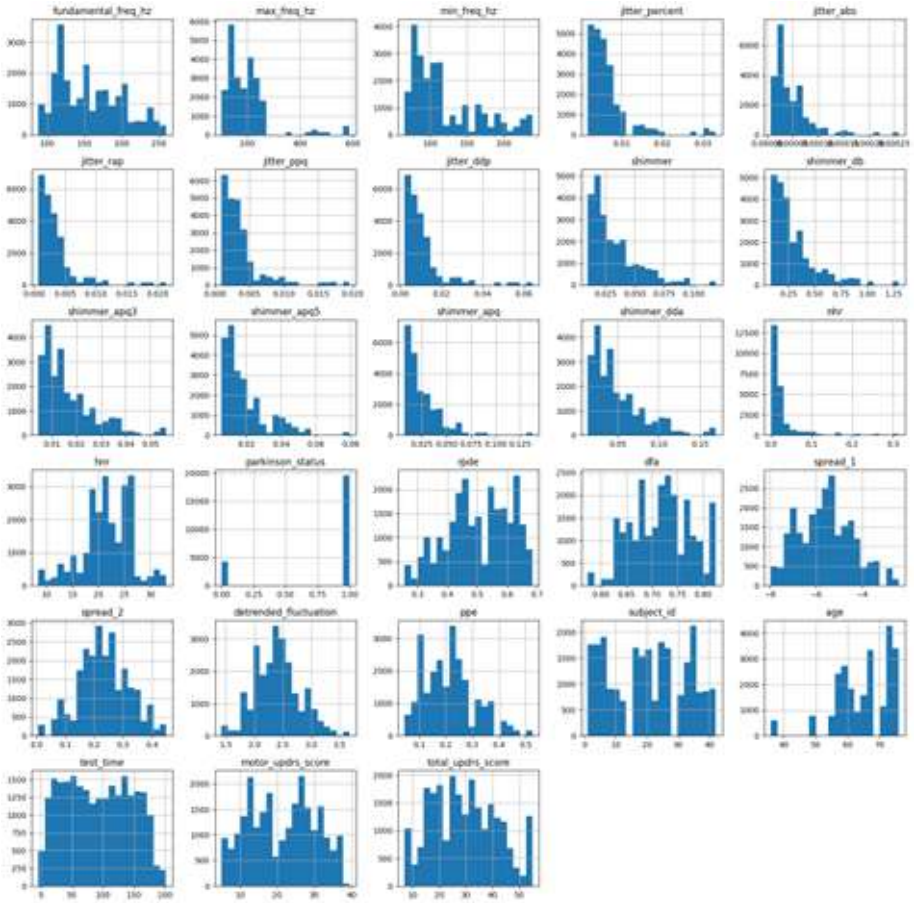

Fig. 4 Value ranges of PD Disease

The next step in selecting the "best" options is to conduct a more thorough examination of the primary subjects. There is solid groundwork for all of the key themes. We removed areas with low contrast and unstable edges. Afterwards, each remaining critical point is given an orientation. This method relies on near-pixel gradient orientations. The values are weighted according to the gradient magnitudes. With the generated points, feature vectors—sometimes called descriptors—are built. This computation makes use of the 16x16 neighbourhood that is immediately surrounding the pixel. The directions and magnitudes of the gradients in the neighbourhood may be calculated. There is a Gaussian distribution applied to the weights of these values. Under these conditions, the orientation histograms are generated independently for every sub-region. An output vector of 128 values (16 x 8) is generated by the entire process.

D. Traditional Classification

A wide variety of traditional categorization methods have been utilized for the purpose of diagnosing Parkinson's disease. These methods make use of a variety of characteristics that have been gathered from medical literature. Clinical examinations, imaging studies, and demographic information are all included in these characteristics. A total of five conventional classification approaches, including SVM, DT, kNN, LR and NB classifiers, have been extensively utilized in this work.

E. Integrated Ensembled Techniques

Ensemble categorization is often called "learning by groups." It improves classification system dependability. The ensemble model [24] can combine learners that are barely better than a random guess into strong learners that can make accurate predictions. Ensemble classifiers generalize better than base learners. Ensemble classifiers are typically supervised learning algorithms because they may be learnt from existing data and applied to new data. Ensembles consist of generating basis learning and integrating base learners. An algorithm like a decision tree, neural network, or other machine learning technique uses a base learner created from training data. In picture recognition, medical diagnosis, and illness classification, ensemble classifier techniques have increased performance. This article focuses on classic ensemble techniques: voting, bagging, boosting, and stack generalization. Ensemble classifiers are typically used to improve PD disease diagnostic performance and prediction.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

This ensemble model was built using Python and the Anaconda IDE. It uses AdaBoost, Bagging, soft voting, and weighted voting. Max Little of Oxford University utilized the model to analyze data from the PD illness dataset available in the UCI Machine Learning Repository. Of the 195 biomedical sounds included in the speech signal collection, 48 are for healthy individuals and 147 are for those with Parkinson's disease (PD). There is a clear demarcation between the datasets used for training and testing. Training the ensemble model's sample features is what the training stages are all about. Using the training datasets, you will attempt to forecast the result of test data during the testing phases.

A. Results and Discussions

An evaluation has been performed on the efficacy of basic classifiers, including SVM, DT, LR, k-NN, and NB classifiers within this domain. The average values of accuracy, precision, recall, and F1-score for each base classifier were assessed and shown in Table 1 and Figure 5. The SVM models achieved the highest performance accuracy, attaining 91.2%. The DT method with k-NN model attained the second highest performance level, achieving an accuracy rate of around 88%. The naïve Bayes classifier had the lowest classification performance among the five fundamental classifiers, attaining an accuracy rate of 81.2%.

Table 1 Performance Analysis of Base Classifiers

S. No.	Base Classifiers	Accuracy	Precision	Recall	F1-Score
1.	SVM	91.5	90.7	90.6	90.7
2.	Decision Tree	89.4	89.1	89.5	89.3
3.	kNN	87.1	86	86	86
4.	Logistic Regression	85.7	85.6	85.1	85.3
5.	Naive Bayes	83.9	83.7	83	83.5

Furthermore, it is essential to include individual or fundamental classifiers with other classifiers in order to maximize the effectiveness of these classifiers. Within the context of our collaborative efforts, we have successfully articulated a collection of five distinct ensemble learning models. Within the framework of the SVM classifier, the bagging approach has been utilized. Within the framework of the SVM classifier, the Adaboost methodology has been included. In the context of the weighted voting method, we have utilized an ensemble approach that is comprised of three fundamental classifiers. These classifiers are the DT, the kNN, and the SVM. In the end, the approach of majority voting makes use of a collection of fundamental classifiers, such as the DT, the kNN, and the SVM.

The performance of several ensembling techniques on the base models is presented in Table 2 and Figure 6. These strategies include SVM-Linear, decision tree, kNN, Logistic Regression, and naïve bayes. The data shown in the table indicates that the ensemble models have a higher level of performance, even though the individual base classifiers have a somewhat lower level of performance. In addition, it was discovered that the implementation of the majority voting technique by utilizing decision trees, MLPs, and SVM-Linears resulted in the highest accuracy of 95.3%.

On the other hand, the bagging method that was utilized in conjunction with Random Forest produced the lowest accuracy amounting to 91.25%. On the other hand, it outperforms each of the five fundamental classifiers that were investigated.

Table 2 Performance Analysis of Ensemble Classifiers

S. No	Integrated Classifiers	Acc.	Prec.	Rec.	F1-Sc.
1.	SVM + AdaBoost	93.15	93.1	93.4	93.2
2.	SVM + Bagging	92.25	91.75	91.5	91.6
3.	kNN+ DT + SVM + MV	93.50	94	94	93
4.	kNN + DT + SVM + WV	95.3	95.2	95.4	95.3

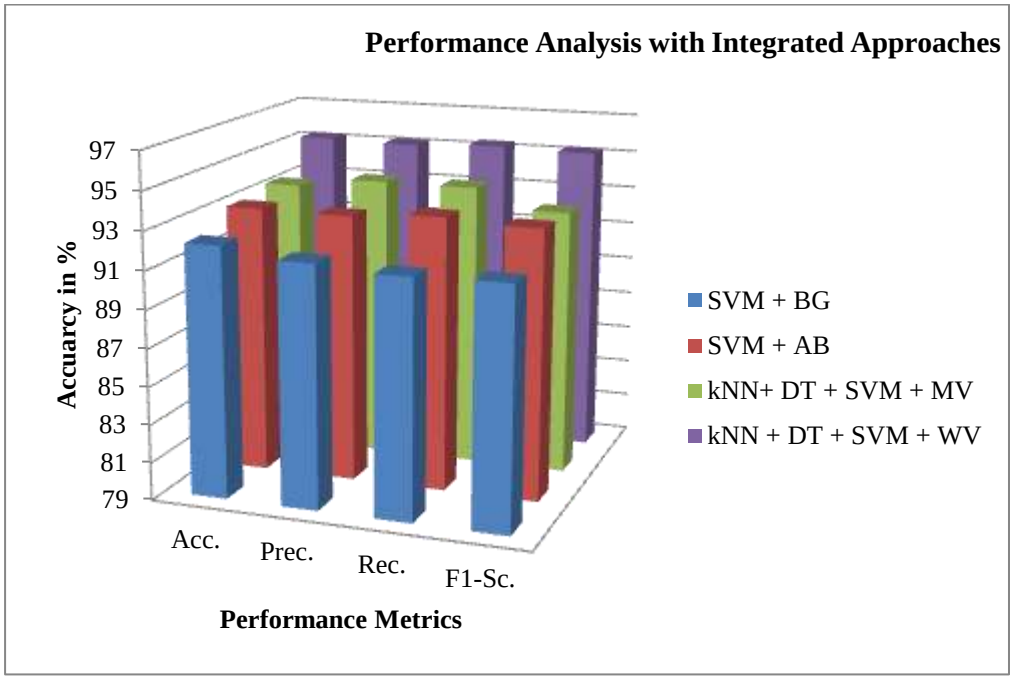

Fig. 6 Performance comparison with proposed ensembled approaches

V.CONCLUSIONS

Identifying and predicting Parkinson's disease through speech signal analysis presents significant challenges due to the complexity and variability of frequency patterns. This study demonstrates that ensemble learning, particularly the majority voting approach, enhances predictive performance compared to traditional single-model classifiers. Among the ensemble methods evaluated—Adaboost, Bagging, Majority Voting, and Weighted Voting—majority voting achieved the highest classification accuracy, while Adaboost underperformed relative to the others. Despite this, ensemble models outperformed strong individual classifiers such as Support Vector Machines and Naïve Bayes, indicating the robustness of ensemble strategies in this domain. Future research will focus on integrating deep learning and ensemble deep learning frameworks to better handle large-scale datasets. Additionally, efforts will be directed toward optimizing training time through GPU acceleration, thereby improving scalability and computational efficiency for real-time applications.

REFERENCES

[1]. Rizek, P.; Kumar, N.; Jog, M.S. An update on the diagnosis and treatment of Parkinson disease. CMAJ 2016, 188, 1157–1165.

[2]. Karan, B.; Sahu, S.S.; Mahto, K. Parkinson disease prediction using intrinsic mode function based features from speech signal. Biocybern. Biomed. Eng. 2019, 40, 249–264.

[3]. Rawat, A.S.; Rana, A.; Kumar, A.; Bagwari, A. Application of multi layer artificial neural network in the diagnosis system: A systematic review. IAES Int. J. Artif. Intell. 2018, 7, 138.

[4]. Available online: https://www.who.int/news-room/fact-sheets/detail/parkinson-disease (accessed on 30 October 2022).

[5]. K. Kalaivani, M. Priya, P. Deepan, L.R. Sudha, and J. Ganesh, "Heart Disease Prediction System Based on Multiple Feature Selection Algorithm with Ensemble Classifier", ECS Transactions, Vol. 107 (1), pp. 8049-8059, 2022. https://doi.org/10.1149/10701.8049ecst.

[6]. Tolosa, E.; Garrido, A.; Scholz, S.W.; Poewe, W. Challenges in the diagnosis of Parkinson's disease. Lancet Neurol. 2021, 20, 385–397.

[7]. Blauwendraat, C.; Nalls, M.A.; Singleton, A.B. The genetic architecture of Parkinson's disease. Lancet Neurol. 2020, 19, 170–178.

[8]. Armstrong, M.J.; Okun, M.S. Diagnosis and Treatment of Parkinson Disease: A Review. JAMA 2020, 323, pp. 548–560.

[9]. Deepan, P., Vidhya, R., Rajalingam, B., Santhoshkumar, R., Arul, N. (2024). FLAML-HDPS Model: An Efficient and Intelligent AutoML Approach for Heart Disease Prediction. In: Devi, B.R., Kumar, K., Raju, M., Raju, K.S., Sellathurai, M. (eds) Proceedings of Fifth International Conference on Computer

and Communication Technologies. IC3T 2023. Lecture Notes in Networks and Systems, vol 897. Springer, Singapore. https://doi.org/10.1007/978-981-99-9704-6_25

[10]. Vijiaratnam, N.; Simuni, T.; Bandmann, O.; Morris, H.R.; Foltynie, T. Progress towards therapies for disease modification in Parkinson's disease. Lancet Neurol. 2021, Vol. 20, pp. 559–572.

[11]. Rana, A.; Rawat, A.S.; Bijalwan, A.; Bahuguna, H. Application of multi-layer (perceptron) artificial neural network in the diagnosis system: A systematic review. In Proceedings of the 2018 International Conference on Research in Intelligent and Computing in Engineering (RICE), San Salvador, El Salvador, 22–24 August 2018; pp. 1–6.

[12]. Simon, D.; Tanner, C.; Brundin, P. Parkinson Disease Epidemiology, Pathology, Genetics, and Pathophysiology. Clin. Geriatr. Med. 2020, 36, 1–12.

[13]. Dong-Chen, X.; Yong, C.; Yang, X.; Chen-Yu, S.; Li-Hua, P. Signaling pathways in Parkinson's disease: Molecular mechanisms and therapeutic interventions. Signal Transduct. Target Ther. 2023, 8, 73.

[14]. Hazan, H.; Hilu, D.; Manevitz, L.; Ramig, L.O.; Sapir, S. Early diagnosis of Parkinson's disease via machine learning on speech data. In Proceedings of the 2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel, Eilat, Israel 14–17 November 2012; pp. 1–4. [CrossRef]

[15]. Al-Sarem, M., Saeed, F., Boulila, W., Emara, A.H., Al-Mohaimeed, M., Errais, M.: Feature selection and classification using CatBoost method for improving the performance of predicting Parkinson's disease. Advances on Smart and Soft Computing. AISC, vol. 1188, pp. 189–199. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-6048-4 17

[16]. Wroge, T.J., ¨Ozkanca, Y., Demiroglu, C., Si, D., Atkins, D.C., Ghomi, R.H.: Parkinson's disease diagnosis using machine learning and voice. In: 2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), pp. 1–7. IEEE (2018)

[17]. Wan, K.R., Maszczyk, T., See, A.A.Q., Dauwels, J., King, N.K.K.: A review on microelectrode recording selection of features for machine learning in deep brain stimulation surgery for Parkinson's disease. Clin. Neurophysiol. 130(1), 145–154 (2019)

[18]. Cavallo, F., Moschetti, A., Esposito, D., Maremmani, C., Rovini, E.: Upper limb motor pre-clinical assessment in Parkinson's disease using machine learning. Parkinsonism Related Disord. 63, 111–116 (2019)

[19]. Almeida, J.S., et al.: Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques. Pattern Recogn. Lett. 125, 55–62 (2019)

[20]. Parisi, L., RaviChandran, N., Manaog, M.L.: Feature-driven machine learning to improve early diagnosis of Parkinson's disease. Expert Syst. Appl. 110, 182–190 (2018)

[21]. Kotsavasiloglou, C., Kostikis, N., Hristu-Varsakelis, D., Arnaoutoglou, M.: Machine learning-based classification of simple drawing movements in Parkinson's disease. Biomed. Signal Process. Control 31, 174–180 (2017)

[22]. Lichman, M. UCI Machine Learning Repository; University of California, School of Information and Computer Science: Irvine, CA, USA; Available online: http://archive.ics.uci.edu/ml (accessed on 25 September 2022).

[23]. David G. Lowe. Object recognition from local scale-invariant features. In International Conference on Computer Vision, 1999.

[24]. P.Deepan and L.R. Sudha, "Scene Classification of Remotely Sensed Images using Ensembled Machine Learning Models", Proceedings in Lecturer Notes on Electrical Engineering, Springer Nature, pp.535-550, 2021, https://doi.org/10.1007/978-981-16-0289-4_39

Corresponding author: G. Harichandana

[1](Student, Dept. of. Computer Science and Engineering, Sri Mittapalli College of Engineering,Guntur, India)