IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Enhancing Parkinson's Disease Diagnosis through Stacking Ensemble-Based Machine Learning Approach

FATMA A. HASHIM[1,2], RIYADH M. AL-TAM[3], SARMAD MAQSOOD[4], LAITH ABUALIGAH[5,6,7,8], REEM M. ALWHAIBI[9]

[1]Faculty of Engineering, Helwan University, Egypt
[2]MEU Research Unit, Middle East University, Amman 11831, Jordan
[3]School of Computational Sciences, Swami Ramanand Teerth Marathwada University, Nanded 431606, Maharashtra, India
[4]Centre of Real Time Computer Systems, Faculty of Informatics, Kaunas University of Technology, LT-51386 Kaunas, Lithuania
[5]Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman 19328, Jordan
[6]Computer Science Department, Al al-Bayt University, Mafraq 25113, Jordan
[7]MEU Research Unit, Middle East University, Amman 11831, Jordan
[8]Applied science research center, Applied science private university, Amman 11931, Jordan
[9]Department of Rehabilitation Sciences, College of Health and Rehabilitation Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia

Corresponding author: F. A. Hashim (fatma_hashim@h-eng.helwan.edu.eg)

This work was supported by the Princess Nourah bint Abdulrahman University Researchers' Supporting Project number (PNURSP2024R117), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

ABSTRACT Parkinson's disease is a progressive neurological condition that affects motor abilities. Common symptoms include tremors, muscle stiffness, and difficulty with coordinated movements. A variety of efforts are underway to address these issues and improve diagnostic precision in Parkinson's disease. This paper employs well-known machine-learning techniques to improve diagnostic accuracy. A variety of individual and ensemble AI models have been proposed, including Random Forest, Decision Tree, Logistic Regression, Gradient Boosting, Support Vector Machine, Stacking, and Bagging Ensemble classifiers. Three scenarios are applied to two standard benchmark datasets. The best performance is achieved when the Stacking Ensemble classifier is utilized, where the Support Vector Machine and Gradient Boosting are engaged for extracting features and Logistic Regression for classifying Parkinson's disease. 1.00,0.00,0.00The Stacking Ensemble classifier reaches 94.87% accuracy and 90.00% AUC for the first dataset, while for the second dataset, 96.18% accuracy and 96.27% AUC are recorded. The final results demonstrate the importance of the suggested framework, which can help to improve the overall diagnosis outcomes.

INDEX TERMS Parkinson's disease, diagnosis, machine learning, classification.

I. INTRODUCTION

Parkinson's disease (PD) impacts approximately 2-3% of individuals below the age of 65, ranking as the second most prevalent neurodegenerative disorder characterized by progressive deterioration [1]. Because of the disease's spread, this degeneration starts in the dorsal striatum and moves toward the ventral area [2]. The striatum, comprised of the putamen and caudate nucleus, regulates a spectrum of motor and cognitive functions. In PD, elevated levels of reactive oxygen species generated during dopamine metabolism result in increased iron content, potentially damaging cellular components and impeding neuronal function [3]. As of now, no definitive cure for PD exists, with available interventions limited to surgical procedures and medication, albeit accompanied by side effects impacting individuals' daily lives [4]. The depletion of dopaminergic neurons, central to PD pathology, triggers a myriad of motor and non-motor symptoms [5]. Tremors, stiffness, slow movement, and difficulty walking are examples of motor symptoms; accidents, depression, psychosis, genitourinary problems, constipation, and sleep

disorders are examples of non-motor symptoms. PD significantly disrupts routine movements and automatic actions, affecting unconscious gestures like smiling or blinking [6]. These symptoms, which appear after 60% of dopaminergic neurons are damaged, are correlated with aging factors [7] and lead to a lower quality of life overall. Clinical diagnosis categorizes PD into five stages, with stages 1 and 2 representing milder forms that allow patients to maintain daily functionality. However, those in stages 4 and 5, unable to move independently, necessitate care from others. The early diagnosis of PD relies on the Hoehn and Yahr scale or the Unified PD Rating Scale (UPDRS) [8]. Yet, challenges persist, as patients self-assess on these scales, introducing inconsistencies and subjectivity. Efforts are ongoing to address these issues and enhance diagnostic precision in the realm of Parkinson's disease.

According to data provided by the World Health Organization (WHO), PD has impacted around 10 million individuals globally [9]. The likelihood of developing PD increases with age, posing a significant concern as it affects 1% of the overall older population. Regrettably, a considerable number of patients do not receive timely diagnosis during the initial stages of the disease, resulting in the development of a persistent neurological condition that is incurable in later phases, often leading to fatalities [10]. About 6.2 million people worldwide were impacted by PD in 2015 alone, which resulted in 117,400 fatalities [11]. Detecting PD poses challenges, as few of its symptoms resemble those observed in cool temperature, such as voice tremors and unsteady movements. Consequently, there is a pressing demand to introduce a method competent of extricating essential features crucial for PD detection [12]. Compounding the challenge is the expense and limited accuracy of current diagnostic tests for the disease. These disconcerting realities underscore the immediate requirement for a cost-effective, effective and precise early-stage diagnostic technique for PD [13]. Such a method would enable timely intervention and potential curative measures before the disease progresses to an incurable state, aligning with ethical considerations and professional standards in healthcare [14].

Currently, a definitive method for diagnosing PD remains elusive. Identifying the disease in its early phases holds the potential for effective eradication through appropriate medication [15]. In clinical practice, physicians rely on a mix of signs and tests for diagnosis to ascertain the presence of PD [16]. Researchers have actively explored various biomarkers as potential indicators for early PD detection, aiming to impede the progression of the disease. While existing therapies can ameliorate PD symptoms, they do not possess the capacity to halt or slow down the disease's advancement. Studies indicate that PD may initiate prior to the onset of motor symptoms, with approximately 90% of PD patients experiencing voice disorders [17]. Consequently, there is a concerted effort to discover more effective means of identifying non-motor symptoms that manifest earlier, offering the prospect of delaying disease progression. However, relying solely on qualitative criteria for PD diagnosis presents challenges, given the potential for other diseases to exhibit similar symptoms. In this context, the consideration of execution time and algorithm complexity becomes crucial, particularly in the realm of medical applications and image analysis [18]–[21].

The landscape of medical image analysis has undergone a transformative shift with the advent of Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) methods [22]. The application of DL extends to a diverse array of activities, including shape modeling, lesion detection, illness categorization, segmentation, and registration [23]. DL approaches, distinguished by their remarkable generalization capacity, are particularly adept at extracting high-level features that enhance accuracy in disease classification. The evolution of Convolutional Neural Networks (CNNs) stands out as a pivotal development propelling the field of medical image analysis forward. CNNs have demonstrated significant efficacy in various medical imaging applications, contributing to notable advancements in the discipline [24] [16]. This progression aligns with ethical considerations and professional standards, as it strives to enhance the accuracy and reliability of medical image analysis, ultimately benefiting patient care and diagnosis.

The dataset related to PD exhibits imbalances within its classes, presenting challenges that can be effectively mitigated through various sampling methods. These methods encompass random oversampling, undersampling, and the application of Synthetic Minority Over-sampling Technique (SMOTE), as well as the utilization of ensemble models [25]. The imbalance in class distribution poses a significant hurdle in enabling the model to efficiently learn the decision boundary, particularly when cases of the minority class are insufficient. Addressing this concern involves exploring oversampling, a method where instances of the minority class are duplicated in the training data before model fitting. While this can balance the distribution of classes, it doesn't introduce new information to the model, raising considerations about the richness of the dataset [26]. Conversely, undersampling aims to balance the intersections between members of the minority class and the majority class by reducing the dataset size. However, this process may result in the loss of some information, potentially posing challenges for DL models in their subsequent training [27].

Typically, PD detection relies on the analysis of speech signals, speech data, or other input modalities. However, existing approaches for detecting and classifying PD using various input data have shown suboptimal performance. Clinical techniques for PD detection predominantly involve labor-intensive, laboratory-based measurements, and computerized methods. Consequently, there exists a pressing need to develop an improved PD detection approaches that enhances classification performance.

Traditionally, the detection of PD involves a comprehensive examination of the neurological background of the patient and a study of their motions in various contexts. Diag-

nosing PD poses inherent challenges, particularly when it's first developing and its engine signs are mild, as there lacks a reliable laboratory test. Patients are routinely required to visit clinics for ongoing assessments to monitor the progression of the disease over time. Recognizing the distinct vocal features present in PD patients, voice recordings emerge as a non-invasive and effective diagnostic tool. Our proposed method exhibits a high level of accuracy in detecting PD while maintaining cost-effectiveness. Notably, it excels in providing early detection, a critical factor in significantly enhancing an individual's quality of life. In contrast to other approaches heavily reliant on ML models analyzing inputs from sensor devices, our approach surpasses them. This not only enhances accuracy but also ensures efficiency and cost-effectiveness in comparison to prevailing algorithms. The significant contributions of this study are outlined as follows:

1) A novel Stacking Ensemble-based approach combines Support Vector Machine, Gradient Boosting, and Logistic Regression to automatically distinguish between healthy individuals and those with PD. Support Vector Machine and Gradient Boosting are employed for feature extraction, while Logistic Regression is utilized for classification.
2) A comprehensive diagnosis has been conducted to effectively and reliably classify Parkinson's disease based on three scenarios: Individual AI model, Bagging Ensemble model, and Stacking Ensemble model classifications.
3) Five individual AI models and the Bagging Ensemble are used for a performance comparison study with the proposed Stacking Ensemble model.
4) 1.00,0.00,0.00The bootstrapping technique is employed to check deeply the achieved results of the proposed AI models.
5) The SMOTE approach is employed for augmentation processing to address over-fitting and create a balanced dataset for the training set.

The remaining sections of this manuscript are systematically organized as follows: Section II outlines notable existing works related to the subject. Section III provides the proposed methodology. Section IV engages in discussions on the experimental results and comparisons with other methods. Section V concludes this work with future research direction.

II. RELATED WORK

Numerous researchers have investigated the application of DL methods for the detection and diagnosis of PD, as documented in the works of various authors [28]–[30]. Diagnosis methods encompass the analysis of diverse data modalities, including voice recordings, brain scan images, and drawings such as meander patterns, spirals, waves, among others [31]. The utilization of DL has become prevalent in the medical imaging field due to its notable accuracy in early-stage PD detection, establishing it as a common tool for predicting PD.

Nilashi et al. [8] introduced a remote tracking system employing a clustering method to predict PD based on voice data. The study utilized the UCI dataset comprising 5875 instances to assess the model's performance. The presented methodology advocated the successful use of clustering, employing Self-Organizing Maps (SOM) to group information determined by similarity. These agglomeration, generated by SOM, were then utilized by artificial neural networks (ANNs) for classification. Subsequently, the similar clusters underwent learning in the next phase, involving a deep neural network (DNN). To evaluate the model's efficacy, the researchers employed the root mean square error score, achieving a commendable score of 0.537 on the test data.

Das et al. [32] conducted a comprehensive analysis comparing various approaches to PD diagnosis. The study aimed to proficiently distinguish healthy individuals through the implementation of four classification patterns, namely Neural Networks (NNs), Regression, DMneural, and Decision Tree (DT). Rigorous comparative research methodologies were employed, encompassing diverse assessment approaches to gauge the performance of these patterns. The findings of the study underscored the superior classification outcomes of NNs compared to Regression, DMneural, and Decision Tree, exhibiting an impressive 92.9% accuracy. Furthermore, the study compared these outcomes with the results obtained from Kernel Support Vector Machine (KSVM), revealing encouraging findings.

Rastegar et al. [33] introduced a method designed for PD detection utilizing RF applied to cytokine data. Cytokine molecular data serves as a valuable source of information pertaining to clinical phenotypes, playing a crucial role in immune system signaling. The authors employed RF for classifying a dataset consisting of records from 360 individuals, sourced from the Michael J Fox Foundation. The tree-like formation inherent to the RF facilitated the detection of PD, leveraging entropy and information derived from the provided information. The system performance was rigorously evaluated, employing the root mean square error (RMSE) metric, resulting in values of 0.1123 for the Hoehn and Yahr scale and 0.1193 for the Unified PD Rating Scale part three (UPDRS III).

Zhao et al. [34] employed a DL method, integrating CNN and Long Short-Term Memory (LSTM) approaches, to analyze gait data for the identification of PD. The gait signals were meticulously modified to ensure accurate transmission to the CNN network. This study included a thorough comparison of the presented system with alternative models and prior research, revealing outstanding outcomes in terms of accuracy and other pertinent measures. In a parallel development, vocal analysis methods have garnered substantial attention from researchers interested in constructing predictive telediagnosis and telemonitoring networks for detecting PD. The researchers leveraged abundant voice signal data sources, collected during conversational activities with both healthy peoples and those diagnosed with PD.

Rehman et al. [35] meticulously gathered data from a

cohort of 31 male and female patients, encompassing a total of 195 voice recordings. The study addresses the concern of imbalanced dataset through the implementation of three sampling approaches to enhance model performance and mitigate overfitting. Experimental outcomes showcase that, with a balanced dataset using random oversampling, the proposed approach achieves impeccable results with 100% accuracy. Additionally, employing the SMOTE yields a model with 91% F1 score.

Sharma et al. [36] proposed an approach for detecting PD on the Unified PD Scale through the utilization of voice data. The method introduced employs a SVM for the detection of PD on this scale. The dataset for this model encompasses 197 instances collected from the UCI inventory. Prior to the regression analysis, the researchers conducted meticulous preprocessing steps, including the application of basic statistical measures to assess the data's mean, median, and null values, aiming to address potential skewness issues. These measures were essential for refining the data quality, since skewness or the existence of null values can significantly impact the generalization of the model. Recognizing the paramount importance of data cleanliness in influencing learning model performance, the researchers systematically removed null data by imputing the mean and improved data skewness through standardized approaches. The resultant preprocessed data underwent regression analysis using a SVM. The model yielded a notable achievement with a 0.24 RMSE.

Chen et al. [37] introduced a DL model designed for the prediction of PD utilizing patient voice data. Within the proposed model, the careful consideration of factors such as both the quantity of neurons and the choice of activation functions proves pivotal in ensuring the accurate classification of data. Remarkably, this meticulous attention to architectural details results in an impressive R2 score of approximately 96% on the training data, showcasing the model's efficacy in capturing and representing the underlying patterns in the voice data associated with PD.

Mahmood et al. [38] proposed a DL based model designed to detect PD. This model demonstrates an impressive level of accuracy, detecting PD with an error of only 0.10 RMSE, surpassing the performance of existing models. Notably, the proposed model excels in extracting essential voice features, facilitating the detection of PD in its early phases.

Little et al. [39] proposed a classification method for distinguishing individuals with PD from control subjects based on dysphonia. They collected data from 31 people, including 23 with PD and 8 healthy persons, yielding 195 sustained vowel-phonations. They used pitch period entropy as a valid dysphonia indicator. The methodology encompassed three key steps: Preprocessing, feature selection, feature calculation, and linear kernel classification. Authors exhibited a commendable accuracy level of 91.4%. In a related effort, Quan et al. [40] utilized DL-based methods for PD identification, conducting a comparison with and without optimization methods. Their utilization of k-fold cross-validation contributed to enhanced accuracy levels. Yasar et al. [41] employed ANN for PD detection, employing a dataset sourced from the UCI repository. The work incorporated one output for categorization out of 45 input properties. The presented model demonstrated a remarkable accuracy of 94.93% in effectively identifying PD patients from healthy participants.

Li et al. [42] developed a hybrid CNN-LSTM model to predict PD from voice signals. The utilization of CNN facilitated the extraction of crucial data and LSTM was instrumental in making predictions. The presented hybrid methodology demonstrated superior performance compared to single-model approaches. In a related study, Ma et al. [43] aimed to detect PD utilizing DL techniques, using the PD dataset for feature extraction and dataset balance. This study achieved an impressive accuracy of 97% in identifying PD.

Within the existing literature, numerous models and frameworks grounded in ML and DL approaches have been developed to detect PD through the analysis of patients voice, gait, and handwriting. These three modalities hold particular significance in PD detection, given the observable alterations in voice, gait movement, and handwriting that accompany the onset of the disease. Such changes can manifest across the entire body, affecting both physical and mental health. In the literature, the detection of PD through voice signals has been a primary focus due to the availability of benchmark datasets and the attainment of the highest accuracy. However, even the study with the lowest RMSE fails to address the multivariate characteristics of the database and lacks factor regarding the significance of features crucial for effective disease detection. The performance of existing single models suggests their limitations in achieving accurate results compared to group DL models together for illness detection. Additionally, the reported outcomes for PD detection exhibit a level of efficacy that warrants further research. In response to these considerations, we propose a novel DL-based hybrid model integrated with sampling methods. This approach aims to address imbalances in dataset classes, enhance generalization performance, and overall enhance the accuracy of PD detection.

III. PROPOSED METHODOLOGY

PD is a neurological condition that develops over time and is most often known for its effects on motor abilities. Through non-invasive speech analysis, this study aims to build a dependable tool for early identification of Parkinson's disease by applying cutting-edge AI models. However, a few steps must be fulfilled before training and testing any model, including data preprocessing, splitting, and augmenting, as shown in Figure 1. To classify the collected data into healthy and PD, a number of individual and ensemble AI models have been suggested, including Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Gradient Boosting (GB), SVM, Stacking, and Bagging Ensemble classifiers.

A. DATA ACQUISITION

In this work, two dataset are used [44], where identifying dysphonia was applied to differentiate between healthy

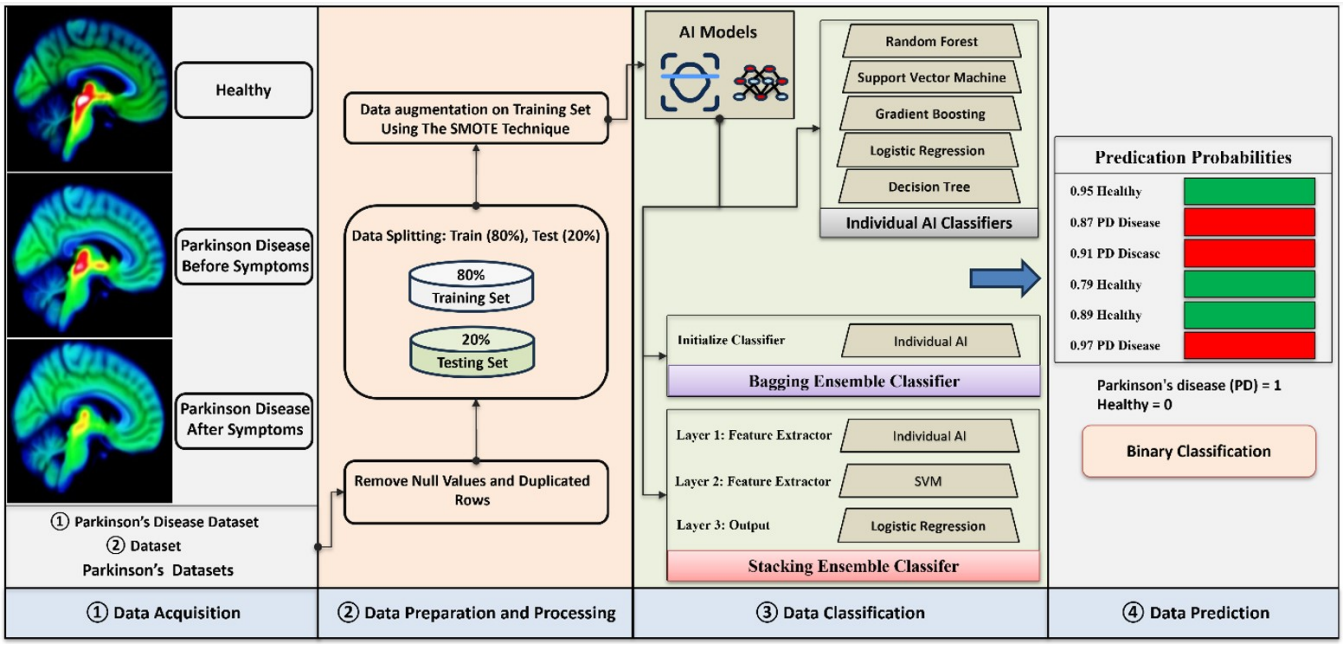

FIGURE 1: The proposed AI models for recognizing Parkinson's disease from healthy ones using data-only datasets.

individuals and those suffering from PD. The first dataset holds biological speech measures that were collected from 31 individuals, some of which had PD [44]. Such a dataset holds 195 voice recordings made by these people, and each column represents a specific voice measure. Moreover, the "status" column has two values: 0 to represent healthy and 1 to denote PD. On the other hand, the second dataset obtained from the Department of Neurology at the CerrahpaÃ©Å¸a Faculty of Medicine, Istanbul University, consisted of 188 PD patients, aged 33 to 87 (81 women and 10 men) [45]. Additionally, 64 healthy people were included in the dataset (41 women and 23 men), aged 41 to 82. The data was collected using a microphone set to 44.1 KHz and resulted in 756 rows and 754 columns of features.

B. DATA PREPARATION AND PROCESSING

The first step, known as preprocessing, is crucial in readying the collected data for classification as healthy or having PD. The collected datasets are typically in a data format that includes specific columns for labeling each row and are read using NumPy's Data Frame for processing. During this process, null values are replaced with zeros, and any duplicated rows are eliminated. 1.00,0.00,0.00In this context, redundant extracted features with more than 30% missing data are eliminated [46]. To address multicollinearity, Spearman correlations between features in pairs are calculated [46], [47]. If the absolute correlation coefficient exceeds 0.8, one of the features is removed. Finally, missing data in the remaining features is imputed using the median values of the corresponding features. Generally, such techniques allow us to include valuable features in our study.

C. DATA SPLITTING

Typically, datasets are divided into two groups (training and testing sets) or three groups (training, validation, testing sets), to train and validate the performance of an AI model. In this study, the used datasets are divided into 80% training and 20% testing sets for binary classification purposes, as shown in Table 1.

TABLE 1: Data description for training (80%) and testing (20%)

Dataset	Data Splitting	Healthy	PD
Dataset 1	Training	38	118
	Testing	10	29
Total		48	147
Dataset 2	Training	292	312
	Testing	74	78
Total		366	390

D. DATA AUGMENTATION

Data augmentation is a method used to increase the size of a dataset by creating a balanced dataset, with the goal of improving overall classification accuracy. However, as mentioned above section, the suggested datasets often have an unbalanced distribution of healthy and PD cases. This is why the SMOTE is used to synthesize new data from the existing

ones [48]. However, only the SMOTE method is applied to the training set after splitting to avoid any overlapping bias.

E. THE PROPOSED DEEP LEARNING MODELS

In this study, individual and ensemble models are used to classify the PD based on data features only. The individual models are RF, DT, LR, GB, and SVM. The random state is set 0 for all individual models to produce the same results every time we train and test the proposed AI models. On the other hand, the bagging and stacking ensemble are built using these individual models.

1) AI-based Individual Random Forest Model

A random forest is a type of machine learning model that uses multiple decision trees to increase prediction accuracy and avoid over-fitting [49]. This is done by training each decision tree on a different subset of the dataset, and then averaging the results.

2) AI-based Individual Decision Tree Model

Decision Trees (DTs) are a type of machine learning algorithm used for both classification and regression tasks [49]. The technique is non-parametric, meaning it does not rely on any assumptions about the underlying distribution of the data. Instead, the algorithm builds a model by recursively partitioning the data into smaller and smaller subsets based on the values of different input features. Each partition corresponds to a basic decision rule, which is used to predict the value of a target variable. In essence, a decision tree is a piecewise constant approximation of the target variable based on the available input features.

3) AI-based Individual Logistic regression Model

Logistic regression is another ML algorithm used for classification tasks [49]. Its goal is to predict the probability that an instance belongs to a specific class or not. Logistic regression analyzes the relationship between two data factors. Logistic regression yields a probability score between 0 and 1 based on the values of the predictor variables, indicating the likelihood of the positive or negative class.

4) AI-based Individual Gradient Boosting Model

Gradient boosting is a widely used technique in machine learning for solving regression and classification problems [49]. It involves an ensemble learning method where the model is trained incrementally, with each new model trying to improve on the previous one. This approach can improve the accuracy of models that are not performing well. AdaBoost and Gradient Boosting are two of the most commonly used boosting algorithms.

5) AI-based Individual SVM Model

Support Vector Machines (SVMs) are powerful machine learning algorithms that can be used for tasks such as regression, classification (including text and image classification, handwriting recognition, spam detection, face detection, and anomaly detection), and outlier identification [49]. SVMs are versatile and effective since they can handle high-dimensional data and nonlinear connections. They work by identifying the largest possible separation hyperplane between the many classes present in the target feature.

6) AI-based Stacking Ensemble Model

The stacking classifier is a technique in which individual models are combined to create a complete model for classification. In stacked generalization, each model's output is stacked, and a classifier is used to make the final prediction [49]. This allows each model to contribute its unique strengths to the final prediction. In this study, each AI individual model is used as the first layer of the proposed stacking classifier for feature extraction. The SVM classifier is used as the second layer for feature extraction, and Logistic Regression is used for the final classification prediction.

7) AI-based Bagging Ensemble Model

The bagging classifier is an ensemble meta-estimator that can independently fits basic classifiers on random subsets of the original dataset, then combines their individual predictions to make a final prediction through voting or averaging [49]. This model requires an initial classifier, so the suggested individual AI models are involved with the goal of reaching the best performance.

F. EXPERIMENTAL SETUP

1.00,0.00,0.00In this work, the selected datasets are exploited to estimate the efficacy and accuracy of the proposed approach for the detection of PD. The training and testing of the classifiers were conducted using a 80:20 split ratio of labeled data. The hyperparameters for the training phase, including the learning rate, epoch, and minibatch size, were initialized as 0.0001, 30, and 64, respectively. The experiments were performed using a 5-fold cross validation.

G. EVALUATION METRICS

1.00,0.00,0.00Many standard evaluation matrices, including as precision, F1-score, sensitivity (Se), accuracy (Acc), and receiver operating characteristics (ROC) curve, are employed for PD classification. The mathematical definitions for each are as follows:

$$Precision(Pre) = \frac{TP}{TP + FP}, \tag{1}$$

$$Recall/Se = \frac{TP}{TP + FN}, \tag{2}$$

$$Acc = \frac{TP + TN}{TP + TN + FN + FP}, \tag{3}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{4}$$

Where FN (False-Negative) indicates a technique incorrectly classifying the disease as negative, TP (True-Positive)

TABLE 2: Experimental evaluation study (%) for the chosen individual AI classifiers using Dataset 1.

AI Model	Class	FP	Accuracy	AUC	Precision	Recall	F1-score
RF	Healthy	4	89.74	80.00	100.00	60.00	75.00
	PD	0			88.00	100.00	94.00
DT	Healthy	4	87.18	78.28	86.00	60.00	71.00
	PD	1			88.00	97.00	92.00
LR	Healthy	3	89.74	83.28	88.00	70.00	78.00
	PD	1			90.00	97.00	93.00
GB	Healthy	4	87.18	78.28	86.00	60.00	71.00
	PD	1			88.00	97.00	92.00
SVM	Healthy	6	84.62	70.00	100.00	40.00	57.00
	PD	0			83.00	100.00	91.00

indicates a method correctly identifying the condition as positive. A technique correctly categorizing an illness as negative is known as TN (True-Negative). Finally, False-Positive (FP) shows a technique incorrectly flagging the illness as positive. Moreover, the ability of a classifier to distinguish between healthy and PD cases is evaluated using the ROC curve with its AUC value [49].

IV. RESULTS AND DISCUSSION

This section encompasses the experimentation and validation of the proposed method. It further incorporates a comprehensive assessment of the evaluation metrics and a comparative analysis with other established approaches, accompanied by a detailed discussion.

A. RESULTS

In this study, many experiments have been conducted through three classification scenarios (A, B, C), aiming to reach the best performance. First, five individual AI models are chosen to classify Parkinson's data into healthy or PD, namely RF, DT, LR, GB, and SVM classifiers. In the second scenario, all individual AI models are used as the initial classifier for the Bagging Ensemble classifier. Finally, three individual AI models have been chosen carefully based on their final results to form a list of classifiers for the Stacking Ensemble classifier. All experiments are conducted using the same parameter configurations, with 100 epochs and a random state parameter value of 0 being set.

1) Scenario A: Parkinson Classification-based AI Individual Classifier

In this section, the final classification performances have been investigated and analyzed among five classifiers, including RF, DT, LR, GB, and SVM. It is shown in Table 2 that the average accuracy and AUC for RF, DT, LR, GB, and SVM are (89.74%, 80.00%), (87.18%, 78.28%), (89.74%, 83.28%), (87.18%, 78.28%), and (84.62%, 70.00%) when the Dataset 1 is used.

On the other hand, confusion matrices are depicted for each model, aiming to show the number of cases that are correctly or incorrectly predicted. Figure 2 shows that RF, DT, LR, GB, and SVM made incorrect predictions, where 4 cases for RF, 5 for DT, 4 for LR, 5 for GB, and 6 cases for SVM, are presented when Dataset 1 is used.

Similarly, as tabulated in Table 3, when Dataset 2 is used, average accuracy and AUC of (94.08%, 94.09%), (82.24%, 82.31%), (94.08%, 94.06%), (94.08%, 94.06%), and (91.45%, 91.35%) for RF, DT, LR, GB, and SVM are achieved, respectively.

Likewise, when Dataset 2 is utilized, 9 cases for RF, 27 for DT, 9 for LR, 9 for GB, and 13 cases for SVM are wrongly predicted, as shown in Figure 3.

1.00,0.00,0.00To deeply check the achieved results, the bootstrapping technique is used to create 5 random samples from the original dataset. The bootstrapping technique is used to rigorously evaluate the performance of the proposed AI models. This technique involves repeatedly resampling the data to create multiple training and testing sets, allowing for a more comprehensive assessment of the models' generalizability and robustness [7]. This technique is only applied to the best results, namely when Dataset 2 was used, as shown in Table 3. As shown in Table 4, the best-recorded results are gained by GB, reaching 94.73% accuracy and 94.79% AUC.

2) Scenario B: Parkinson Classification-based AI Bagging Ensemble Classifier

In the second scenario, each AI model acts as an initial classifier for the Bagging Ensemble model. We found that the performance remains consistent for most classifiers, except for the DT and GB classifiers when using Dataset 1. In this case, the accuracy and AUC values reach 89.74% and 83.28% for both classifiers, respectively, as shown in Table 5. We also generated confusion matrices for the Bagging Ensemble with the proposed AI individual models, which are depicted in Figure 4.

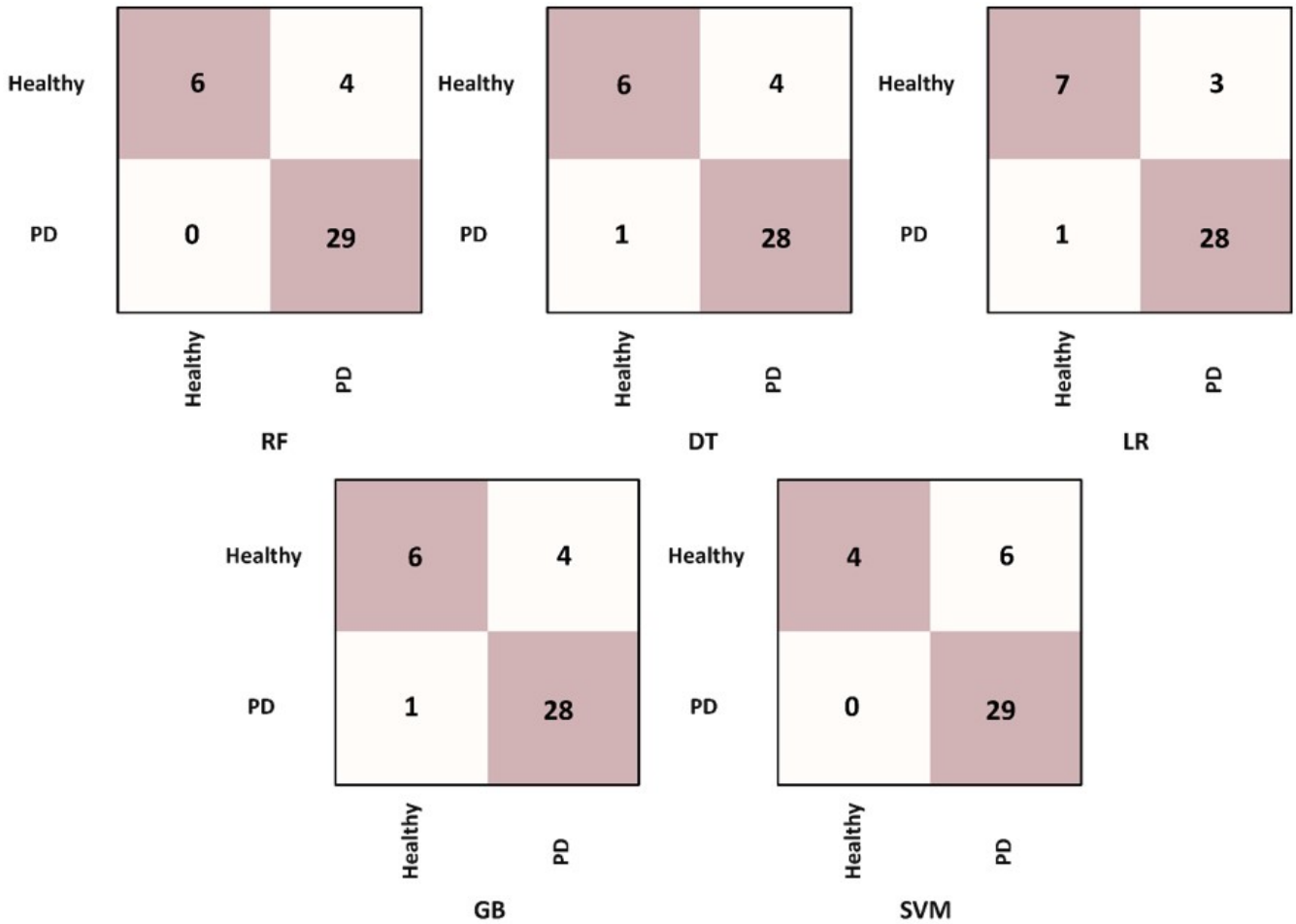

FIGURE 2: The confusion matrices of the individual AI models using Dataset 1.

On the contrary, when using Dataset 2, the performance is improved except for the Bagging with FR classifier which recorded lower results of 92.76% and 92.74% for accuracy and AUC. In this context, the Bagging-based DT classifier reaches 93.42% accuracy and 93.42% AUC, while 95.39% accuracy and 95.41% AUC are gained by using the Bagging-based LR. Besides, (94.74%, 94.73%) and (92.11%, 92.01%) are achieved when the Bagging-based GB and Bagging-based SVM are applied, respectively, as shown in Table 6.

Moreover, as shown in Figure 5, when Dataset 2 is used, six healthy cases are wrongly predicted as a disease and five cases as healthy cases when the Bagging-based RF is involved. Besides, 10, 7, 8, and 12 cases are gotten when the Bagging-based DT, LR, GB, and SVM are applied, respectively.

In contrast, the original dataset is bootstrapped to evaluate the proposed AI models listed in Table 6. Among the initial individual AI models, GB and LR achieve the highest performance, with accuracies of 94.60% and 94.07% and AUCs of 94.42% and 94.09%, respectively, as shown in Table 7.

3) Scenario C: Parkinson Classification-based AI Stacking Ensemble Classifier

In the last scenario, the Stacking Ensemble model is created based on a list of classifiers to extract features and a final classifier to predict cases. Therefore, two configurations are applied: a classifier for feature extraction and a classifier for classification, and two classifiers for feature extraction and a classifier for classification. As shown in scenarios A and B above, LR outperforms all other models; therefore, it is used as a final prediction layer in the Stacking Ensemble model. Furthermore, to determine which individual model can be used for the first or second layers in the second configuration option, a sample procedure is added to select one individual model for the first layer and another or maybe the same for the second layer, as shown in Figure 6. Similarly, the first configuration has only one loop for feature extraction; therefore, the first loop is removed. When the simple procedure is executed on Datasets 1 and 2, many options are generated; however, the second configuration option with three classifiers (SVM for the first layer, GB for the second layer, and

TABLE 3: Experimental evaluation study (%) for the chosen individual AI classifiers using Dataset 2.

AI Model	Class	FP	Accuracy	AUC	Precision	Recall	F1-score
RF	Healthy	4	94.08	94.09	93.00	95.00	94.00
	PD	5			95.00	94.00	94.00
DT	Healthy	11	82.24	82.31	80.00	85.00	82.00
	PD	16			85.00	79.00	82.00
LR	Healthy	5	94.08	94.06	95.00	93.00	94.00
	PD	4			94.00	95.00	94.00
GB	Healthy	5	94.08	94.06	95.00	93.00	94.00
	PD	4			94.00	95.00	94.00
SVM	Healthy	9	91.45	91.35	94.00	88.00	91.00
	PD	4			89.00	95.00	92.00

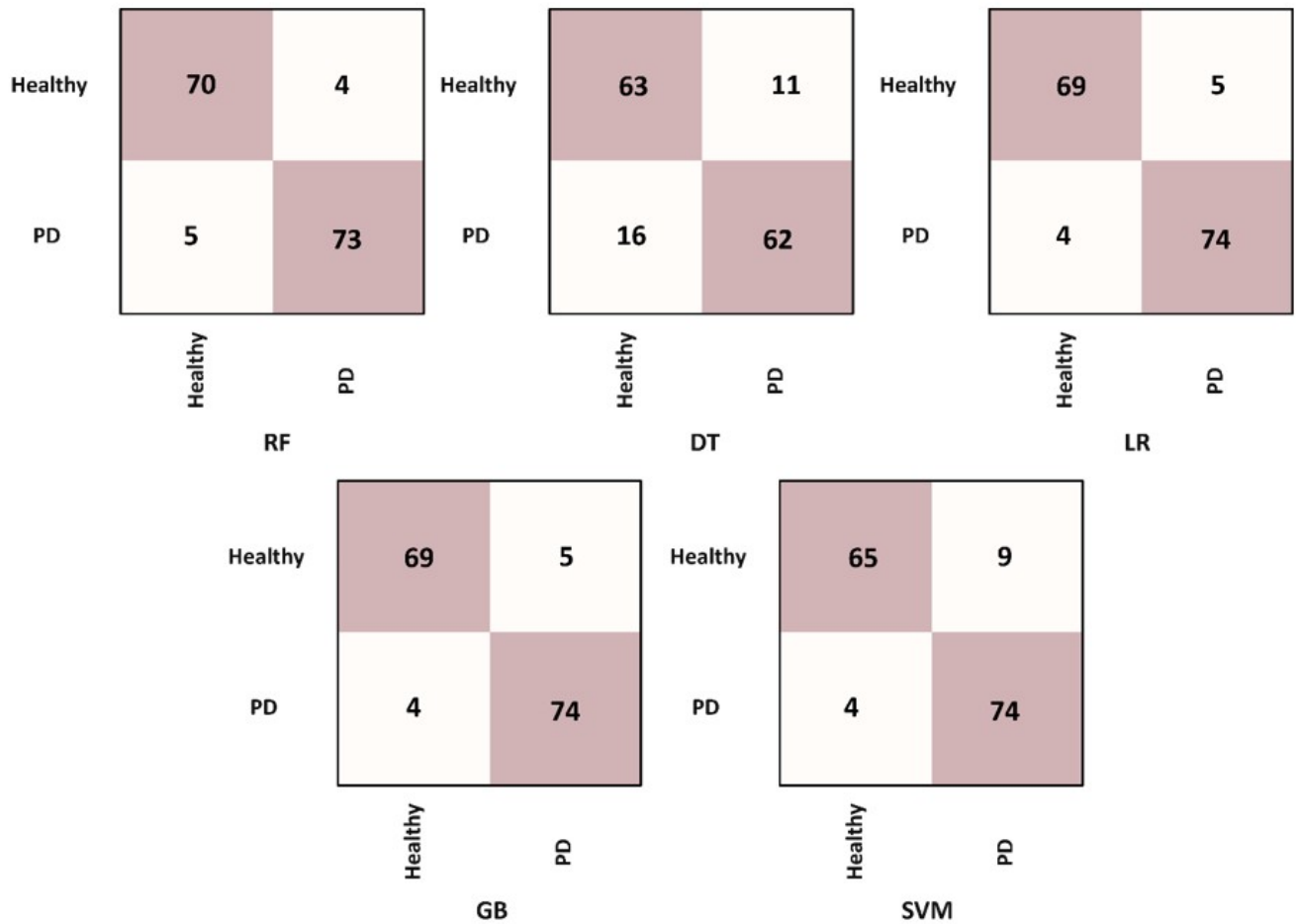

FIGURE 3: The confusion matrices of the individual AI models using Dataset 2.

LR for the final prediction) records the best performance.

When applying the Stacking Ensemble-based SVM, GB, and LR classifiers to Dataset 1, the best results are achieved with an average accuracy of 94.87% and AUC of 90.00%. However, Figure 8 presents a sample of the generated alternatives, showing 92.31% accuracy and 88.28% AUC for both Stacking with RF+GB+LR and LR+GB+LR. Additionally, Stacking with DT+GB+LR and GB+GB+LR reach 92.31%

TABLE 4: 1.00,0.00,0.00Experimental evaluation study (%) for the chosen individual AI classifiers when the Bootstrapping technique is applied on Dataset 2.

AI Model	Sample No.	FP	Accuracy	AUC	Precision	Recall	F1-score
DT	1	(11,6)	88.82	88.82	88.98	88.82	88.80
	2	(8,6)	90.73	90.77	90.78	90.73	90.74
	3	(17,14)	79.47	79.07	79.44	79.47	79.42
	4	(9,11)	86.75	86.77	86.75	86.75	88.77
	5	(15,7)	85.43	83.06	85.74	85.43	85.34
	Total AVG.		**86.24**	**86.09**	**86.33**	**86.24**	**86.61**
LR	1	(4,1)	96.71	96.56	96.88	96.56	96.68
	2	(6,3)	94.07	94.21	93.95	94.21	94.04
	3	(7,3)	93.42	93.66	93.21	93.66	93.32
	4	(1,6)	95.39	95.83	94.99	95.83	95.30
	5	(5,7)	92.10	92.13	92.10	92.13	92.10
	Total AVG.		**94.33**	**94.47**	**94.22**	**94.47**	**94.28**
RF	1	(6,6)	92.11	92.11	92.11	92.11	92.11
	2	(2,0)	98.68	98.78	98.71	98.68	98.68
	3	(5,7)	92.05	92.11	92.10	92.05	92.06
	4	(5,5)	93.38	93.33	93.38	93.38	93.38
	5	(4,2)	96.03	95.93	96.05	96.03	96.02
	Total AVG.		**94.45**	**94.45**	**94.47**	**94.45**	**94.45**
GB	1	(7,0)	95.39	95.73	95.45	95.73	95.38
	2	(8,0)	94.73	94.93	95.06	94.93	94.73
	3	(4,7)	92.76	92.63	92.88	92.63	92.72
	4	(5,1)	96.05	95.77	96.35	95.77	95.99
	5	(0,8)	94.73	94.93	95.06	94.93	94.73
	Total AVG.		**94.73**	**94.79**	**94.96**	**94.79**	**94.71**
SVM	1	(9,10)	87.50	87.50	87.51	87.50	87.50
	2	(7,4)	92.72	92.83	92.81	92.72	92.72
	3	(14,13)	82.12	81.87	82.10	82.12	82.11
	4	(15,12)	82.12	81.81	82.11	82.12	82.08
	5	(7,6)	91.39	91.32	91.39	91.39	91.39
	Total AVG.		**87.17**	**87.06**	**87.18**	**87.17**	**87.16**

accuracy and 85.00% AUC. Finally, confusion matrices are created for these models, as depicted in Figure 7, where the proposed stacking-based SVM+GB+LR incorrectly predicts two cases, while the rest have three incorrect cases.

On the other hand, when Dataset 2 is utilized, the stacking-based SVM+GB+LR model achieves the highest performance, with an accuracy and AUC value of 96.05%, as indicated in Table 9. Additionally, six incorrect cases are identified in Figure 8. The remaining results of other models are noteworthy, with an average accuracy of 94.74% and AUC of 94.73%, while eight incorrect cases are shown.

1.00,0.00,0.00The proposed stacking-based SVM+GB+LR model, as shown in Table 10, outperforms the state-of-the-art performance with an accuracy of 96.18% and an AUC of 96.27%. This superior performance is achieved through the application of the bootstrapping technique.

B. DISCUSSION

1) Discussing the Achieved performance by the Proposed AI models

In Scenario A, Tables 2-3 display experiments to determine the optimal individual AI model to classify Parkinson's disease. The LR model outperforms all others. Additionally, confusion matrices for these models show that the LR model has 4 incorrect predictions on Dataset 1 and 9 on Dataset 2. However, the poorest performance is observed when using SVM for Dataset 1, with 6 incorrect predictions, and DT with 27 incorrect predictions for Dataset 2, as depicted in Figures 2-3. In Scenario B, the individual AI models serve as initial classifiers for the Bagging Ensemble model, which is then trained and tested on Dataset 1. Bagging-based RF, LR, and SVM show no change in performance, while the others enhance overall performance, as indicated in Table 5 and Figure 4. However, when Dataset 2 is used, the performance is improved across the board, except for the Bagging-based FR model, which exhibits lower performance

TABLE 5: Experimental evaluation study (%) for the chosen individual AI as initial classifiers for the Bagging Ensemble using Dataset 1.

AI Model	Class	FP	Accuracy	AUC	Precision	Recall	F1-score
RF	Healthy	4	89.74	80.00	100.00	60.00	75.00
	PD	0			88.00	100.00	94.00
DT	Healthy	3	89.74	83.28	88.00	70.00	78.00
	PD	1			90.00	97.00	93.00
LR	Healthy	3	89.74	83.28	88.00	70.00	78.00
	PD	1			90.00	97.00	93.00
GB	Healthy	3	89.74	83.28	88.00	70.00	78.00
	PD	1			90.00	97.00	93.00
SVM	Healthy	6	84.62	70.00	100.00	40.00	57.00
	PD	0			83.00	100.00	91.00

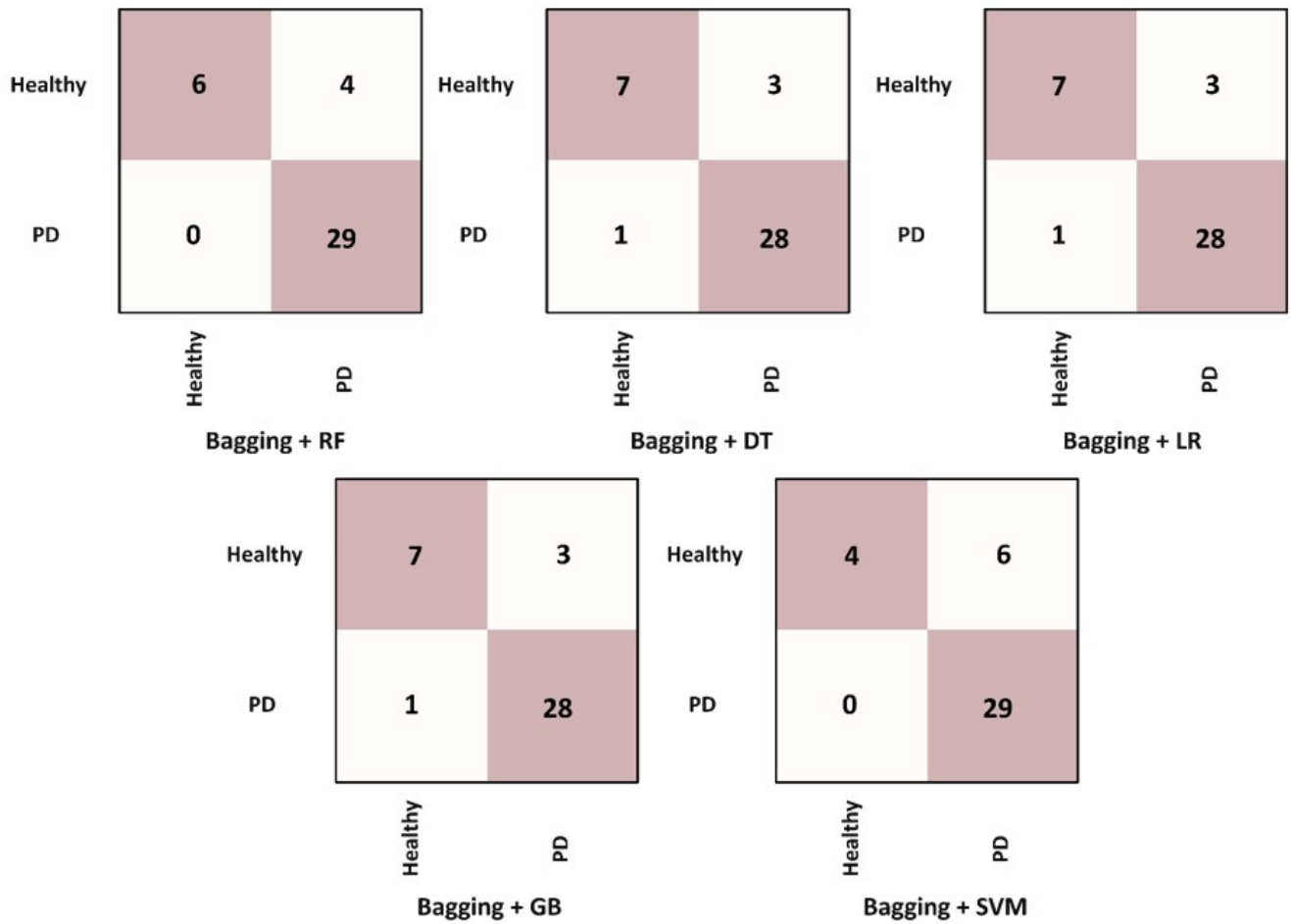

FIGURE 4: The confusion matrices of the Bagging Ensemble models using Dataset 1.

compared to using RF alone, as presented in Table 6 and Figure 5. Finally, in Scenario C, the proposed Stacking-based SVM, GB, and LR, are identified as the best classifiers based on the previous experimental results. The feature extraction models are selected to belong to the Stack Ensemble based on their final results. It is evident that the proposed approach

TABLE 6: Experimental evaluation study (%) for the chosen individual AI as initial classifiers for the Bagging Ensemble using Dataset 2.

AI Model	Class	FP	Accuracy	AUC	Precision	Recall	F1-score
RF	Healthy	6	92.76	92.74	93.00	92.00	93.00
	PD	5			92.00	94.00	93.00
DT	Healthy	5	93.42	93.42	93.00	93.00	93.00
	PD	5			94.00	94.00	94.00
LR	Healthy	3	95.39	95.41	95.00	96.00	95.00
	PD	4			96.00	95.00	96.00
GB	Healthy	4	94.74	94.73	95.00	95.00	95.00
	PD	4			95.00	95.00	95.00
SVM	Healthy	8	92.11	92.01	94.00	89.00	92.00
	PD	4			90.00	95.00	92.00

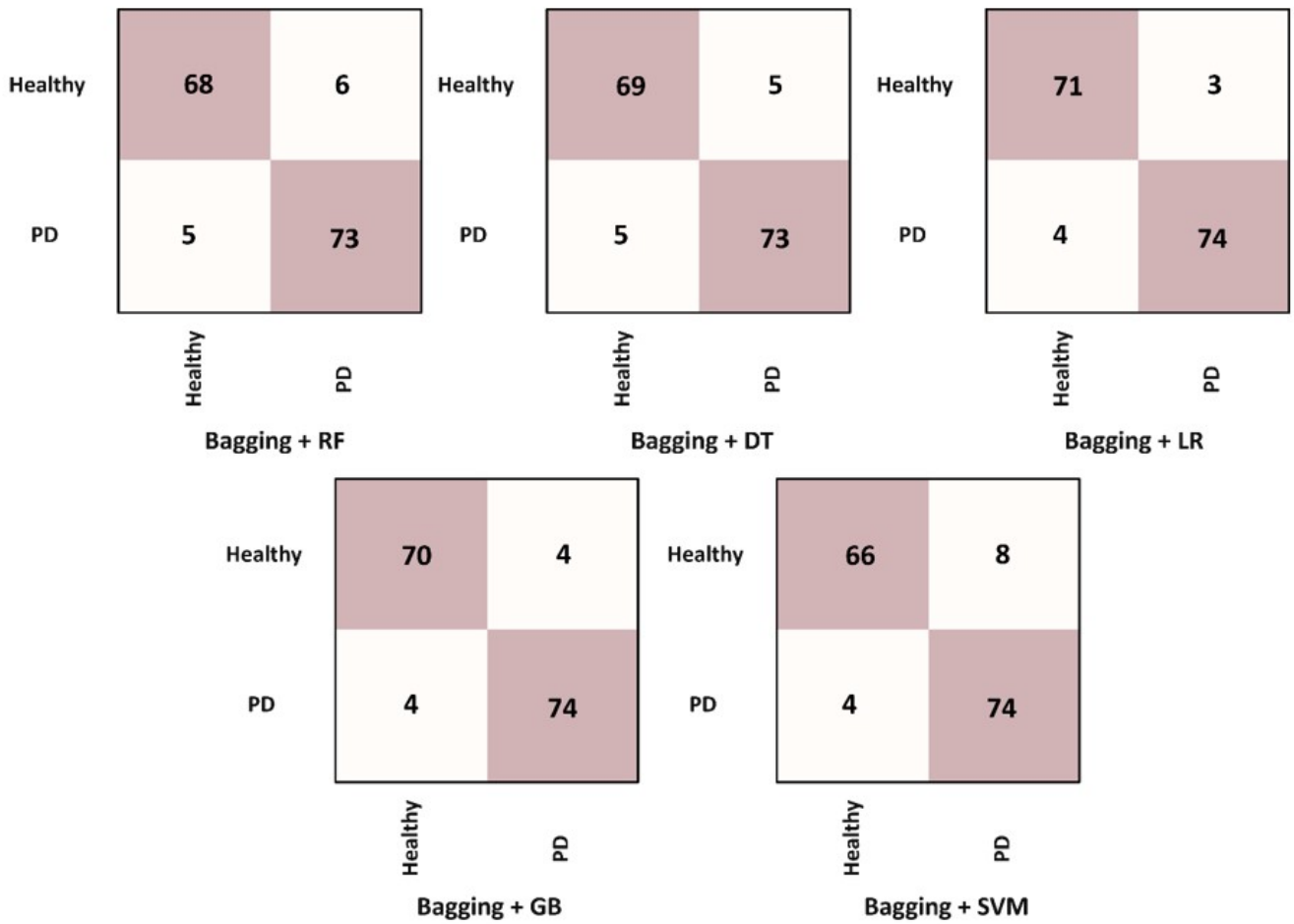

FIGURE 5: The confusion matrices of the Bagging Ensemble models using Dataset 2.

achieves superior evaluation metrics compared to the outcomes of individual AI or Bagging-based models, achieving 94.87% accuracy and 90.00% AUC on Dataset 1, as well as 96.05% accuracy and 96.05% AUC on Dataset 2, as shown

TABLE 7: 1.00,0.00,0.00Experimental evaluation study (%) for the chosen individual AI as initial classifiers for the Bagging Ensemble when the Bootstrapping technique is applied on Dataset 2.

AI Model	Sample No.	FP	Accuracy	AUC	Precision	Recall	F1-score
DT	1	(0,9)	94.07	94.76	94.00	94.76	94.04
	2	(5,10)	90.13	90.40	89.80	90.40	90.00
	3	(2,8)	93.42	93.06	93.8	93.06	93.32
	4	(8,1)	94.07	94.01	94.50	94.01	94.05
	5	(3,10)	91.44	91.61	91.66	91.61	91.44
	Total AVG.		**92.62**	**92.76**	**92.75**	**92.76**	**92.57**
LR	1	(5,6)	93.42	93.35	93.47	93.35	93.40
	2	(6,4)	94.73	94.80	94.67	94.80	94.72
	3	(2,2)	94.07	94.07	93.94	94.07	94.00
	4	(4,1)	94.73	94.82	94.53	94.82	94.6
	5	(5,3)	93.42	93.41	93.41	93.41	93.41
	Total AVG.		**94.07**	**94.09**	**94.00**	**94.09**	**94.02**
RF	1	(6,6)	92.10	92.09	92.09	92.09	92.09
	2	(4,6)	93.42	93.45	93.08	93.45	93.25
	3	(4,3)	95.39	95.27	95.40	95.27	95.33
	4	(9,10)	87.50	87.23	87.07	87.23	87.15
	5	(9,7)	89.47	89.30	89.49	89.30	89.38
	Total AVG.		**91.57**	**91.46**	**91.42**	**91.46**	**91.44**
GB	1	(3,4)	95.39	95.40	95.38	95.40	95.39
	2	(5,5)	93.42	93.41	93.41	93.41	93.41
	3	(1,8)	94.07	93.52	94.78	93.52	93.94
	4	(7,3)	93.42	93.06	93.6	93.06	93.30
	5	(1,4)	96.71	96.73	96.76	96.73	96.71
	Total AVG.		**94.60**	**94.42**	**94.78**	**94.42**	**94.55**
SVM	1	(11,11)	85.52	85.52	85.52	85.52	85.52
	2	(13,7)	86.84	86.08	87.10	86.08	86.44
	3	(8,7)	90.13	89.86	90.01	89.86	89.93
	4	(13,11)	84.21	84.22	84.13	84.22	84.16
	5	(7,11)	88.15	87.87	88.30	87.87	88.02
	Total AVG.		**86.97**	**86.71**	**87.01**	**86.71**	**86.81**

in Tables 8-9 and Figures 6-7. 1.00,0.00,0.00To ensure a robust evaluation, the bootstrapping technique is employed to create five random samples from the original dataset. This approach enhances the reliability of the final evaluation matrices. Notably, bootstrapping is only applied to the best-performing results that were obtained using Dataset 2. As demonstrated in Tables 4, 7, and 10, this strategy generally improves overall performance.

When a large enough dataset is available, the Stacking Ensemble model is the most effective technique to enhance the evaluation matrices. Furthermore, as compared to individual AI or Bagging Ensemble models, the suggested framework can increase the total classification rate; however, the training and testing phases will take more time. Because the devices are always evolving, time is not a key concern when we apply this framework in real applications.

2) A comparison of the related work with the proposed Stacking Ensemble Model

This section compares the assessment findings of the suggested Stacking Ensemble Model with the most recent research on the categorization of Parkinson's disease, as shown in Table 11. It's possible that the suggested framework can provide dependable and motivating findings for use in real classifications. This research establishes an indirect summary comparison with comparable studies that have utilized the proposed datasets of this work. Such a comprehensive comparison with state-of-the-art studies is lacking due to the diverse settings utilized, including different evaluation matrices, different data splitting, different parameter configurations of the created algorithms, or even different dataset distribution.

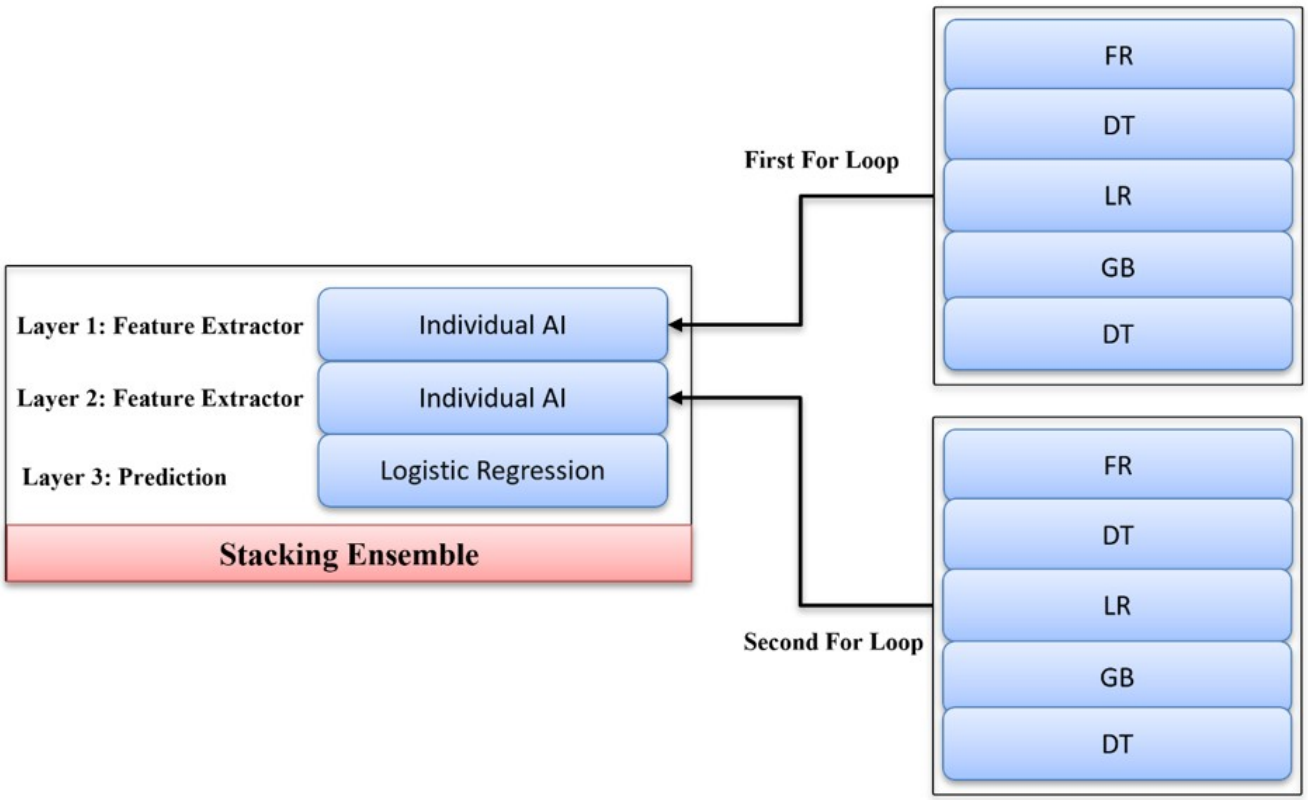

FIGURE 6: A sample procedure to build a list of classifiers for the Stacking Ensemble.

3) Limitation and Future Work

In this work, a limited of machine learning models are used; however, the level would increase if we applied other techniques, including Vision Transformer [50], Ensemble-based concatenating layer [51], or other deep learning models [52]. Furthermore, extra datasets can be used to check the capability of the proposed framework to classify Parkinson's disease.

V. CONCLUSION

Timely diagnosis of PD is crucial for effective patient care. An early diagnosis significantly enhances the likelihood of successful recovery through appropriate treatments and medications. In the contemporary era, computer aided diagnosis have witnessed considerable advancements, particularly in the realms of ML and DL, bringing about a transformative impact on the medical field. These sophisticated approaches have proven instrumental in classifying various diseases, monitoring patient health, and enabling early predictions of medical conditions. In this study, we have utilized different ML models i.e., Random Forest, Decision Tree, Logistic Regression, Gradient Boosting, and SVM including bagging and stacking ensembles to these models and the experimental results shows enhanced results using the stacking ensemble classifier on the selected datasets. 1.00,0.00,0.00The bootstrapping technique is used to create five random samples from the original dataset, in order to achieve the best performance. The proposed method enhance detection accuracy while reducing misinterpretations. The model effectively differentiates between PD and healthy patients, demonstrating outstanding performance accuracy. In future endeavors, we plan to explore more advanced approaches to extract the most crucial features from the dataset for PD detection. Additionally, we will rigorously assess the outcomes using an independent dataset to ascertain the resilience and dependability of the proposed method.

REFERENCES

[1] Kim, J. J., Bandres-Ciga, S., Blauwendraat, C., Gan-Or, Z., & International Parkinson's Disease Genomics Consortium. (2020). No genetic evidence for involvement of alcohol dehydrogenase genes in risk for Parkinson's disease. Neurobiology of aging, 87, 140-e19.

[2] Bernardo, L. S., Quezada, A., Munoz, R., Maia, F. M., Pereira, C. R., Wu, W., & de Albuquerque, V. H. C. (2019). Handwritten pattern recognition for early Parkinson's disease diagnosis. Pattern recognition letters, 125, 78-84.

[3] Dharani, M. K., & Thamilselvan, R. (2023). Hybrid optimization enabled deep learning model for Parkinson's disease classification. The Imaging Science Journal, 1-16.

[4] Rehman, A., Saba, T., Mujahid, M., Alamri, F. S., & ElHakim, N. (2023). Parkinson's disease detection using hybrid lstm-gru deep learning model. Electronics, 12(13), 2856.

[5] Chaudhuri, K. R., & Schapira, A. H. (2009). Non-motor symptoms of Parkinson's disease: dopaminergic pathophysiology and treatment. The Lancet Neurology, 8(5), 464-474.

[6] Poewe, W., Seppi, K., Tanner, C. M., Halliday, G. M., Brundin, P., Volk-

TABLE 8: Experimental evaluation study (%) for the chosen individual AI as initial classifiers for the Bagging Ensemble using Dataset 2.

Stacking Ensemble Model	Class	FP	Accuracy	AUC	Precision	Recall	F1-score
RF+GB+LR	Healthy	2	92.31	88.28	89.00	80.00	84.00
	PD	1			93.00	97.00	95.00
DT+GB+LR	Healthy	3	92.31	85.00	100.00	70.00	82.00
	PD	0			91.00	100.00	95.00
LR+GB+LR	Healthy	2	92.31	88.28	89.00	80.00	84.00
	PD	1			93.00	97.00	95.00
GB+GB+LR	Healthy	3	92.31	85.00	100.00	70.00	82.00
	PD	0			91.00	100.00	95.00
SVM+GB+LR	Healthy	2	94.87	90.00	100.00	80.00	89.00
	PD	0			94.00	100.00	97.00

TABLE 9: A sample of experiments to select the best-fit individual AI for the first layer of the Stacking Ensemble using Dataset 2, where GB is used for the second layer and LR for the predication layer.

Stacking Ensemble Model	Class	FP	Accuracy	AUC	Precision	Recall	F1-score
RF+GB+LR	Healthy	4	94.74	94.73	95.00	95.00	95.00
	PD	4			95.00	95.00	95.00
DT+GB+LR	Healthy	4	94.74	94.73	95.00	95.00	95.00
	PD	4			95.00	95.00	95.00
LR+GB+LR	Healthy	4	94.74	94.73	95.00	95.00	95.00
	PD	4			95.00	95.00	95.00
GB+GB+LR	Healthy	4	94.74	94.73	95.00	95.00	95.00
	PD	4			95.00	95.00	95.00
SVM+GB+LR	Healthy	3	96.05	96.05	96.00	96.00	96.00
	PD	3			96.00	96.00	96.00

mann, J., ... & Lang, A. E. (2017). Parkinson disease. Nature reviews Disease primers, 3(1), 1-21.

[7] Despotovic, V., Skovranek, T., & Schommer, C. (2020). Speech based estimation of Parkinson's disease using Gaussian processes and automatic relevance determination. Neurocomputing, 401, 173-181.

[8] Nilashi, M., Ahmadi, H., Sheikhtaheri, A., Naemi, R., Alotaibi, R., Alarood, A. A., ... & Zhao, J. (2020). Remote tracking of Parkinson's disease progression using ensembles of deep belief network and self-organizing map. Expert Systems with Applications, 159, 113562.

[9] Schiess, N., Cataldi, R., Okun, M. S., Fothergill-Misbah, N., Dorsey, E. R., Bloem, B. R., ... & Dua, T. (2022). Six action steps to address global disparities in Parkinson disease: a World Health Organization priority. JAMA neurology, 79(9), 929-936.

[10] Mohammed, M. A., Elhoseny, M., Abdulkareem, K. H., Mostafa, S. A., & Maashi, M. S. (2021). A multi-agent feature selection and hybrid classification model for Parkinson's disease diagnosis. ACM Transactions on Multimidia Computing Communications and Applications, 17(2s), 1-22.

[11] Pei, X., Fan, H., & Tang, Y. (2021). Temporal pyramid attention-based spatiotemporal fusion model for Parkinson's disease diagnosis from gait data. IET Signal Processing, 15(2), 80-87.

[12] Morris, H. R., Spillantini, M. G., Sue, C. M., & Williams-Gray, C. H. (2024). The pathogenesis of Parkinson's disease. The Lancet, 403(10423), 293-304.

[13] Siciliano, M., Tessitore, A., Morgante, F., Goldman, J. G., & Ricciardi, L. (2024). Subjective Cognitive Complaints in Parkinson's Disease: A Systematic Review and Meta-Analysis. Movement Disorders.

[14] Hoglinger, G. U., Adler, C. H., Berg, D., Klein, C., Outeiro, T. F., Poewe, W., ... & Lang, A. E. (2024). A biological classification of Parkinson's disease: the SynNeurGe research diagnostic criteria. The Lancet Neurology, 23(2), 191-204.

[15] Malek, N., & Grosset, D. G. (2015). Medication adherence in patients with Parkinson's disease. CNS drugs, 29, 47-53.

[16] Chopade, P., Chopade, N., Zhao, Z., Mitragotri, S., Liao, R., & Chandran Suja, V. (2023). Alzheimer's and Parkinson's disease therapies in the clinic. Bioengineering & Translational Medicine, 8(1), e10367.

[17] Sakar, C. O., & Kursun, O. (2010). Telediagnosis of Parkinson's disease using measurements of dysphonia. Journal of medical systems, 34, 591-599.

[18] Ali, N. A., abbassi, A. E., & Cherradi, B. (2022). The performances of iterative type-2 fuzzy C-mean on GPU for image segmentation. The Journal of Supercomputing, 78(2), 1583-1601.

[19] Ait Ali, N., Cherradi, B., El Abbassi, A., Bouattane, O., & Youssfi, M. (2018). GPU fuzzy c-means algorithm implementations: performance analysis on medical image segmentation. Multimedia Tools and Applications, 77(16), 21221-21243.

[20] Maskeliunas, R., Damasevicius, R., Kulikajevas, A., Padervinskis, E., PribuiÅ¡is, K., & Uloza, V. (2022). A hybrid U-lossian deep learning network for screening and evaluating Parkinson's disease. Applied Sciences, 12(22), 11601.

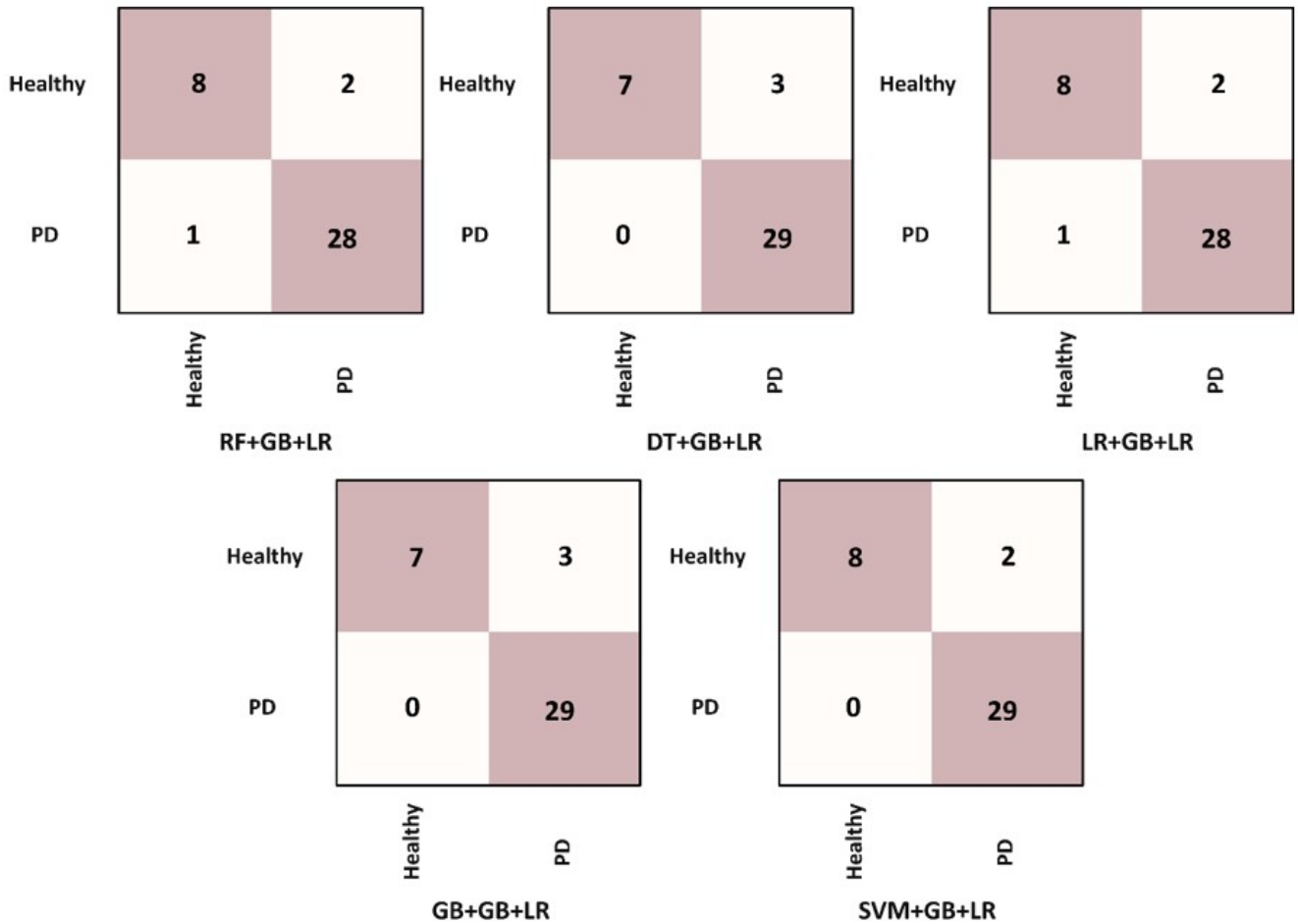

FIGURE 7: The confusion matrices of the Stacking Ensemble models using Dataset 1.

[21] Almeida, J. S., RebouÃ§as Filho, P. P., Carneiro, T., Wei, W., Damasevicius, R., Maskeliunas, R., & de Albuquerque, V. H. C. (2019). Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques. Pattern Recognition Letters, 125, 55-62.

[22] Dolz, J., Desrosiers, C., & Ayed, I. B. (2018). 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. NeuroImage, 170, 456-470.

[23] Priya, S. J., Rani, A. J., Subathra, M. S. P., Mohammed, M. A., Damasevicius, R., & Ubendran, N. (2021). Local pattern transformation based feature extraction for recognition of Parkinson's disease based on gait signals. Diagnostics, 11(8), 1395.

[24] Abayomi-Alli, O. O., Damasevicius, R., Maskeliunas, R., & Abayomi-Alli, A. (2020, September). BiLSTM with data augmentation using interpolation methods to improve early detection of parkinson disease. In 2020 15th conference on computer science and information systems (FedCSIS) (pp. 371-380). IEEE.

[25] Wang, S. H., Phillips, P., Sui, Y., Liu, B., Yang, M., & Cheng, H. (2018). Classification of Alzheimer's disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling. Journal of medical systems, 42, 1-11.

[26] Lauraitis, A., Maskeliunas, R., Damasevicius, R., & Krilavicius, T. (2020). A mobile application for smart computer-aided self-administered testing of cognition, speech, and motor impairment. Sensors, 20(11), 3236.

[27] Sigcha, L., Borzi, L., Amato, F., Rechichi, I., Ramos-Romero, C., Cardenas, A., ... & Olmo, G. (2023). Deep learning and wearable sensors for the diagnosis and monitoring of Parkinson's disease: A systematic review. Expert Systems with Applications, 120541.

[28] Camps, J., Sama, A., Martin, M., Rodriguez-Martin, D., Perez-Lopez, C., Arostegui, J. M. M., ... & Rodriguez-Molinero, A. (2018). Deep learning for freezing of gait detection in Parkinson's disease patients in their homes using a waist-worn inertial measurement unit. Knowledge-Based Systems, 139, 119-131.

[29] Nilashi, M., Ahmadi, H., Manaf, A. A., Rashid, T. A., Samad, S., Shahmoradi, L., ... & Akbari, E. (2020). Coronary heart disease diagnosis through self-organizing map and fuzzy support vector machine with incremental updates. International Journal of Fuzzy Systems, 22, 1376-1388.

[30] Ortiz, A., Murcia, F. J. M., Munilla, J., Gorriz, J. M., & Ramirez, J. (2019). Label aided deep ranking for the automatic diagnosis of Parkinsonian syndromes. Neurocomputing, 330, 162-171.

[31] Fan, S., & Sun, Y. (2022, September). Early Detection of Parkinson's Disease using Machine Learning and Convolutional Neural Networks from Drawing Movements. In CS & IT Conference Proceedings (Vol. 12, No. 15). CS & IT Conference Proceedings.

[32] Das, R. (2010). A comparison of multiple classification methods for diagnosis of Parkinson disease. Expert Systems with Applications, 37(2), 1568-1572.

[33] Ahmadi Rastegar, D., Ho, N., Halliday, G. M., & Dzamko, N. (2019). Parkinson's progression prediction using machine learning and serum cytokines. NPJ Parkinson's disease, 5(1), 14.

[34] Asuroglu, T., & Ogul, H. (2022). A deep learning approach for parkinson's disease severity assessment. Health and Technology, 12(5), 943-953.

[35] Rehman, A., Saba, T., Mujahid, M., Alamri, F. S., & ElHakim, N. (2023). Parkinson's disease detection using hybrid lstm-gru deep learning model. Electronics, 12(13), 2856.

[36] Sharma, R. K., & Gupta, A. K. (2015). Voice analysis for Telediagnosis of Parkinson disease using artificial neural networks and support vector

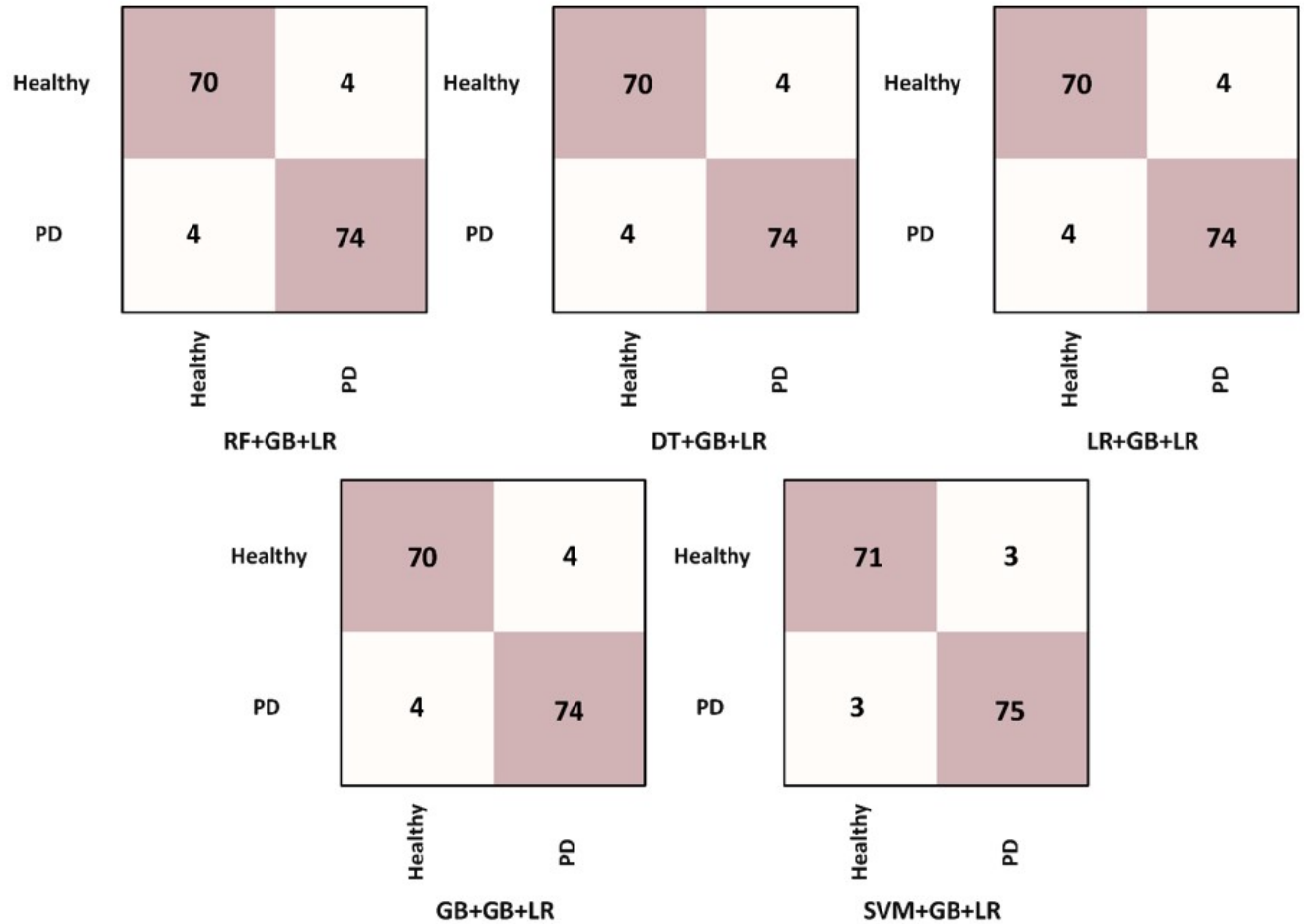

FIGURE 8: The confusion matrices of the Stacking Ensemble models using Dataset 2.

machines. International Journal of Intelligent Systems and Applications, 7(6), 41.

[37] Chen, H. L., Wang, G., Ma, C., Cai, Z. N., Liu, W. B., & Wang, S. J. (2016). An efficient hybrid kernel extreme learning machine approach for early diagnosis of Parkinson's disease. Neurocomputing, 184, 131-144.

[38] Mahmood, A., Mehroz Khan, M., Imran, M., Alhajlah, O., Dhahri, H., & Karamat, T. (2023). End-to-End Deep Learning Method for Detection of Invasive Parkinson's Disease. Diagnostics, 13(6), 1088.

[39] Little, M., McSharry, P., Hunter, E., Spielman, J., & Ramig, L. (2008). Suitability of dysphonia measurements for telemonitoring of Parkinsonâ€™s disease. Nature Precedings, 1-1.

[40] Quan, C., Ren, K., Luo, Z., Chen, Z., & Ling, Y. (2022). End-to-end deep learning approach for Parkinson's disease detection from speech signals. Biocybernetics and Biomedical Engineering, 42(2), 556-574.

[41] Yasar, A., Saritas, I., Sahman, M. A., & Cinar, A. C. (2019, November). Classification of Parkinson disease data with artificial neural networks. In IOP conference series: materials science and engineering (Vol. 675, No. 1, p. 012031). IOP Publishing.

[42] Li, K., Ao, B., Wu, X., Wen, Q., Ul Haq, E., & Yin, J. (2023). Parkinson's disease detection and classification using EEG based on deep CNN-LSTM model. Biotechnology and Genetic Engineering Reviews, 1-20.

[43] Ma, Y. W., Chen, J. L., Chen, Y. J., & Lai, Y. H. (2023). Explainable deep learning architecture for early diagnosis of Parkinson's disease. Soft Computing, 27(5), 2729-2738.

[44] Little, M., McSharry, P., Hunter, E., Spielman, J., & Ramig, L. (2008). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. Nature Precedings, 1-1.

[45] C. Sakar, G. Serbes, A. Gunduz, H. Nizam, and B. Sakar, "Parkinson's Disease Classification," UCI Mach. Learn. Repos., 2018. The UCI Machine Learning Repository.

[46] Wang, T., Paschalidis, A., Liu, Q., Liu, Y., Yuan, Y., & Paschalidis, I. C. (2020). Predictive models of mortality for hospitalized patients with COVID-19: retrospective cohort study. JMIR medical informatics, 8(10), e21788.

[47] Huang, Y. C., Hong, C. T., Chi, W. C., Yen, C. F., Liao, H. F., Liou, T. H., & Chan, L. (2024). Deterioration of Fine Motor Skills and Functional Disability in Patients with Moderate-to-Advanced Parkinson Disease: A Longitudinal Follow-Up Study. Archives of Gerontology and Geriatrics, 105366.

[48] Chawla, N. V. (2010). Data mining for imbalanced datasets: An overview. Data mining and knowledge discovery handbook, 875-886.

[49] Pedregosa, F. (2011). Scikit-learn: Machine learning in python Fabian. Journal of machine learning research, 12, 2825.

[50] Al-Tam, R. M., Al-Hejri, A. M., Narangale, S. M., Samee, N. A., Mahmoud, N. F., Al-Masni, M. A., & Al-Antari, M. A. (2022). A hybrid workflow of residual convolutional transformer encoder for breast cancer classification using digital X-ray mammograms. Biomedicines, 10(11), 2971.

[51] Al-Hejri, A. M., Al-Tam, R. M., Fazea, M., Sable, A. H., Lee, S., & Al-Antari, M. A. (2022). ETECADx: Ensemble self-attention transformer encoder for breast cancer diagnosis using full-field digital X-ray breast images. Diagnostics, 13(1), 89.

[52] Houssein, E. H., Mohamed, O., Samee, N. A., Mahmoud, N. F., Talaat, R., Al-Hejri, A. M., & Al-Tam, R. M. (2023). Using deep DenseNet with cyclical learning rate to classify leukocytes for leukemia identification. Frontiers in Oncology, 13.

TABLE 10: 1.00,0.00,0.00A sample of experiments to select the best-fit individual AI for the first layer of the Stacking Ensemble when the Bootstrapping technique is applied on Dataset 2, where GB is used for the second layer and LR for the predication layer.

AI Model	Sample No.	FP	Accuracy	AUC	Precision	Recall	F1-score
RF+GB+LR	1	(7,8)	90.13	90.13	90.13	90.13	90.13
	2	(3,9)	92.10	92.36	92.10	92.36	92.09
	3	(3,4)	95.39	95.38	95.4	95.38	95.39
	4	(3,1)	97.36	97.49	97.19	97.49	97.33
	5	(2,4)	96.05	95.89	96.15	95.89	96.00
	Total AVG.		**94.20**	**94.25**	**94.19**	**94.25**	**94.18**
DT+GB+LR	1	(3,8)	92.76	92.76	92.94	92.76	92.75
	2	(2,2)	97.36	97.36	97.36	97.36	97.36
	3	(4,4)	94.73	94.70	94.70	94.70	94.70
	4	(5,7)	92.1	92.15	91.95	92.15	92.03
	5	(6,4)	93.42	93.48	93.27	93.48	93.36
	Total AVG.		**94.07**	**94.09**	**94.04**	**94.09**	**94.04**
LR+GB+LR	1	(3,6)	94.07	94.18	94.02	94.18	94.06
	2	(6,6)	92.10	92.07	92.07	92.07	92.07
	3	(3,4)	95.39	95.36	95.41	95.36	95.38
	4	(2,6)	94.73	94.70	94.88	94.70	94.72
	5	(5,4)	94.07	94.09	94.06	94.09	94.07
	Total AVG.		**94.07**	**94.08**	**94.08**	**94.08**	**94.06**
GB+GB+LR	1	(7,0)	95.39	95.7	96.02	95.70	95.33
	2	(4,4)	94.73	94.72	94.72	94.72	94.72
	3	(4,10)	90.78	91.04	90.78	91.04	90.77
	4	(4,0)	97.36	97.50	97.36	97.50	97.36
	5	(2,6)	94.73	94.88	94.70	94.88	94.72
	Total AVG.		**94.59**	**94.76**	**94.71**	**94.76**	**94.58**
SVM+GB+LR	1	(5,1)	96.05	96.33	95.81	96.33	96.00
	2	(3,3)	96.05	96.02	96.02	96.02	96.02
	3	(1,7)	94.73	95.05	94.68	95.05	94.72
	4	(0,6)	96.05	96.20	96.20	96.20	96.05
	5	(3,0)	98.02	97.79	98.27	97.79	97.99
	Total AVG.		**96.18**	**96.27**	**96.19**	**96.27**	**96.15**

FATMA A. HASHIM is an Assistant Professor at Helwan University, Egypt. She received a Ph.D. degree in Biomedical Engineering from Helwan University, Egypt, in 2020. Along with her career, she was a technical reviewer and editorial board member for several international journals. She has more than 20 scientific research papers published in prestigious international journals in the topics of medical imaging processing, bioinformatics, machine learning and its applications. Her research interests include image processing, signal processing, artificial intelligence, bioinformatics, optimization, metaheuristics. She is serving as reviewer for several journals including : knowledge based system, Applied Soft Computing and Engineering Applications of Artificial Intelligence. And She is participating as editor for Applied Soft Computing, Information sciences, Journal of Ophthalmological Diseases, and journal of Electronic research and Application.

RIYADH M. AL-TAM received a B.Sc. degree in Computer Sciences & Information Systems from the Faculty of Computer Sciences & Information Systems, Thamar University. In 2015, he earned an M.Sc. degree in Informatics Engineering from the University of Algarve, Faro, Portugal. He is currently a Ph.D. candidate at the School of Computational Sciences, the Swami Ramanand Teerth Marathwada University, Nanded, India. His main research interests include applications of machine learning in medicine and security.

TABLE 11: Evaluation results of the proposed Stacking Ensemble model for Parkinson's disease classification are compared to the most recent AI research works.

Authors	**Dataset**	**Labels**	**Methodology**	**Accuracy (%)**
Nilashi et al. [8]	UCI dataset	Healthy/PD	SOM+ANN	RMSE=0.546
Das et al. [32]	Max Little dataset	Healthy/PD	Neural Networks, DMneural, Regression, and Decision Tree	Accuracy=92.9%
Rastegar et al. [33]	Michael J Fox dataset	Healthy/PD	ML models	NRMSE=0.1123, Hoehn and Yahr scale=0.1193
Zhao et al. [34]	Physionet Gait in Parkinson's Disease dataset	Healthy/PD	CNN+LSTM	Accuracy=98.7%
Rehman et al. [35]	IAC dataset	Healthy/PD	LSTM+GRU	AUC=99%, F1-score=91%
Sharma et al. [36]	23 features dataset	Healthy/PD	ANN+SVM	Accuracy=96%
Chen et al. [37]	UCI dataset	Healthy/PD	Hybrid ELM+KELM	Accuracy of 96.47%
Mahmood et al. [38]	UCI dataset	Healthy/PD	DL method	RMSE=0.10
Little et al. [39]	Private dataset	Healthy/PD	Pitch Period Entropy (PPE)	Accuracy=91.4%
Quan et al. [40]	PC-PITA dataset	Healthy/PD	2D-CNNs+1D-CNN	Accuracy=92.0%
Yasar et al. [41]	UCI dataset	Healthy/PD	ANN	Accuracy=94.93%
Li et al. [42]	EEG dataset	Healthy/PD	CNN+LSTM	Accuracy=98.60%
The Proposed Stacking	Dataset 1+Dataset 2	Healthy/PD	**Stacking-based SVM+GB+LR**	Dataset 1: 94.87% (Acc.), 90.00% (AUC.) Dataset 2: 96.05% (Acc.), 96.05% (AUC)

SARMAD MAQSOOD received his B.Sc. and M.Sc. degrees in electronic engineering from the International Islamic University, Islamabad, Pakistan, in 2015 and 2018, respectively. He is currently Ph.D. student at the Faculty of Informatics Engineering, Kaunas University of Technology, Lithuania. His research interests include image processing, medical image analysis, multimodal data fusion and deep learning methods. He is the author of more than 15 research articles. He is also a reviewer of many prestigious journals in Elsevier, Springer, ACM, and others.

LAITH ABUALIGAH is an Associate Professor at Computer Science Department, Al Al-Bayt University, Jordan. He is also a distinguished researcher at many prestigious universities. He received a Ph.D. degree from the School of Computer Science at Universiti Sains Malaysia (USM), Malaysia, in 2018. According to the report published by Clarivate, He is one of the Highly Cited Researchers in 2021-2023 and the 1% influential Researchers, which depicts the 6,938 top scientists in the world. In addition, the first researcher in the domain of Computer Science in Jordan for 2021-2023. According to the report published by Stanford University, He is one of the 2% influential scholars, which depicts the 100,000 top scientists in the world. He has published more than 500 journal papers and books, which collectively have been cited more than 19000 times (H-index = 64). His main research interests focus on Artificial Intelligence, Meta-heuristic Modeling, and Optimization Algorithms, Evolutionary Computations, Information Retrieval, Text clustering, Feature Selection, Combinatorial Problems, Optimization, Advanced Machine Learning, Big data, and Natural Language Processing. He currently serves as an associate editor of many prestigious Journals in Elsevier, Springer, ACM, IEEE, and others.

REEM M. ALWHAIBI is a dedicated professional within the Rehabilitation Sciences Department, College of Health and Rehabilitation Sciences at Princess Nourah University, my expertise lies in developmental neurorehabilitation. With a comprehensive background encompassing both clinical and academic realms, I have immersed myself in researching various neurological conditions affecting adults and pediatrics. My commitment to advancing knowledge is evident through the publication of numerous research articles in indexed journals. In addition to my scholarly pursuits, I have undertaken the translation of three significant books from English to Arabic within the field of physical therapy. Beyond academia, my innovative contributions extend to the registration of three unique innovations, reflecting a passion for pushing the boundaries of rehabilitationÂ sciences.