IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 September 2023, accepted 18 October 2023, date of publication 30 October 2023, date of current version 8 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3328331

RESEARCH ARTICLE

Explainable Artificial Intelligence of Multi-Level Stacking Ensemble for Detection of Alzheimer's Disease Based on Particle Swarm Optimization and the Sub-Scores of Cognitive Biomarkers

ABDULAZIZ ALMOHIMEED[1], REDHWAN M. A. SAAD[2], SHERIF MOSTAFA[3], NORA MAHMOUD EL-RASHIDY[4], SARAH FARRAG[5], ABDELKAREEM GABALLAH[6], MOHAMED ABD ELAZIZ[7], SHAKER EL-SAPPAGH[7,8], AND HAGER SALEH[3]

[1]College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic Universitym (IMSIU), Riyadh 11652, Saudi Arabia
[2]College of Informatics, Midocean University, Moroni, Comoros
[3]Faculty of Computers and Artificial Intelligence, South Valley University, Hurghada, Egypt
[4]Machine Learning and Information Retrieval Department, Faculty of Artificial Intelligence, Kafrelsheikh University, Kafrelsheikh 13518, Egypt
[5]Faculty of Computers and Informations, South Valley University, Qena, Egypt
[6]Faculty of Artificial Intelligence, Kafrelsheikh University, Kafrelsheikh 13518, Egypt
[7]Faculty of Computer Science and Engineering, Galala University, Suez 435611, Egypt
[8]Information Systems Department, Faculty of Computers and Artificial Intelligence, Benha University, Banha 13518, Egypt

Corresponding authors: Abdulaziz AlMohimeed (aialmohimeed@imamu.edu.sa) and Hager Saleh (hager.saleh.fci@gmail.com)

This work was supported by Midocean University.

ABSTRACT Alzheimer's disease (AD) is a progressive neurological disorder characterized by memory loss and cognitive decline, affecting millions worldwide. Early detection is crucial for effective treatment, as it can slow disease progression and improve quality of life. Machine learning has shown promise in AD detection using various medical modalities. In this paper, we propose a novel multi-level stacking model that combines heterogeneous models and modalities to predict different classes of AD. The modalities include cognitive sub-scores (e.g., clinical dementia rating – sum of boxes, Alzheimer's disease assessment scale) from the Alzheimer's Disease Neuroimaging Initiative dataset. In the proposed approach, in level 1, we used six base models (Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), K-nearest Neighbors (KNN), and Native Bayes (NB)to train each modality (ADAS, CDR, and FQA). Then, we build stacking training that combines the outputs of each base model for the training set and staking testing that combines the outcomes of each model for the testing set. In level 2, three stacking models are produced for each modality that trains and evaluates based on the output of 6 base models based on (RF, LR, DT, SVM, KNN, and NB) are combined in training stacking for the training set and testing stacking for the testing set. Stacking training is used to train meta-learners (RF), and stacking testing is used to evaluate meta-learners (RF). Finally, in level 3, the output prediction of the stacking model from each modality (ADAS, CDR, and FQA) in the training and testing datasets is merged to build a new dataset, which is staking training and stacking testing. Training stacking is used to train the meta-learner, and the testing set is used to evaluate the meta-learner and produce the final prediction. Our research also aims to provide model explanations, ensuring efficiency, effectiveness, and trust through explainable artificial intelligence (XAI). Feature selection optimization based on Particle Swarm Optimization is used to select the most appropriate sub-scores. The proposed model shows significant potential for improving early disease diagnosis. The results demonstrate that the multi-modality approach outperforms single-modality approaches. Moreover, the proposed multi-level stacking models achieve the highest performance with selected features compared to regular ML classifiers and stacking models using full multi-modalities, achieving accuracy, precision, recall, and F1-scores of 92.08%, 92.07%, 92.08%, and 92.01% for two classes, and 90.03%, 90.19%, 90.03%, and 90.05% for three classes, respectively.

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

INDEX TERMS Machine learning, multi-level stacking machine learning, ensemble learning, sub scores of cognitive, Alzheimer's disease, explainable artificial intelligence, particle swarm optimization.

I. INTRODUCTION

Alzheimer's disease (AD) is the most prevalent form of dementia, characterized by memory loss and cognitive impairment [1], [2]. It is a chronic illness that progressively worsens, leading to severe dementia symptoms [3]. AD causes mild loss of memory impairment in the initial phases, but as the condition worsens, patients become incapable of converse and adapting to their environment. Over time, symptoms significantly disrupt daily activities [3]. Currently, over 50 million individuals worldwide are affected by AD, and this number is projected to triple by 2050 [4]. Unfortunately, there is no known cure for AD, and available treatments only slow down its progression [5]. Mild Cognitive Impairment (MCI) is a condition that may precede AD, but not all individuals with MCI develop AD [6], [7], [8].

Early detection of high-risk patients transitioning from MCI to AD is essential. It has been observed that approximately 10% to 15% of MCI patients convert to AD each year, and after six years, around 80% of MCI patients progress to AD, known as progressive MCI (pMCI). However, some MCI patients, referred to as stable MCI (sMCI), either remain stable or even revert to normal [9]. AD patients have no effective treatments, and present therapies can only slow the illness's course [10]. Therefore, there is a critical need for effective models and biomarkers for the early detection of MCI-C. This makes it possible to treat MCI patients before they develop AD, decreasing the number of AD patients. Effective models and biomarkers for early detection of MCI are urgently needed to reduce the number of AD patients.

Machine learning (ML) techniques have shown promise in predicting AD and monitoring its progression using high-dimensional data [11], [12], [13], [14], [15]. For example In the authors compared different models, SVM, LR, DT, and RF, to predict Alzheimer's disease. The SVM model achieved high performance, the same in [16], [17], and [18], which achieved adequate performance. In addition, ensemble learning employs many algorithms to provide higher detection performance than its base models. It combines diverse algorithms [19], [20]. Stacking combines heterogeneous base learners through a non-deterministic meta-learning algorithm [21]. This approach aims to learn optimal weights for combining the base classifiers. While previous studies have focused on single modalities, such as neuroimaging data from MRI or PET scans [22], [23], [24], [25] it is essential to consider multiple modalities and cognitive scores for accurate AD diagnosis [26]. Medical specialists typically analyze the patient's profile, including various modalities, to improve the diagnosis [27]. Integrating different neuroimaging data has been employed to enhance the performance of AD diagnosis [26]. Furthermore, considering the cost and time associated with gathering neuroimaging data, developing models that leverage multiple modalities is crucial.

Data fusion, the process of integrating information from multiple sources or modalities, is crucial in improving the performance of machine learning (ML models). Data fusion enhances the representation and understanding of complex phenomena by combining diverse data types, such as images [28], text [29], [30], and biological markers [31]. Integrating heterogeneous data sources provides a more comprehensive view and enables the extraction of meaningful patterns and relationships that may not be apparent when considering individual modalities alone. This holistic approach improves ML and DL models' accuracy, robustness, and generalization capabilities [32]. Data fusion significantly impacts various fields, including healthcare [33], [34], and environmental monitoring [33]. In healthcare, for example, combining clinical data, medical images, genetic information, and patient records allows for more accurate disease diagnosis, personalized treatment recommendations, and monitoring of patient outcomes. The fusion of multimodal neuroimaging data, such as positron emission tomography (PET) [24] and magnetic resonance imaging (MRI) [25] which are medical imaging techniques that can detect a phenotype characterized by cortical and hippocampal atrophy [27] MRI, PET, has shown great promise in the early detection and diagnosis of neurological disorders, including Alzheimer's [26].

Accordingly, data fusion is paramount in ML, as it allows for integrating diverse data types and enhances model performance. By combining information from multiple sources or modalities, data fusion improves accuracy, robustness, and generalization capabilities [34]. In ML, data fusion techniques can be categorized into early, late, and hybrid fusion approaches. Early fusion combines raw data from different sources at the input level, enabling joint feature extraction. Late fusion, conversely, merges the outputs of individual models or networks, allowing for the combination of predictions. Hybrid fusion techniques leverage early and late fusion strategies to exploit the complementary information in multiple modalities fully [26], [28].

There are many cost-effective cognitive (neuro-psychological) scores (CSs) administered by a clinical expert to detect AD. These scores include functional activities questionnaire (FAQ), clinical dementia rating – a sum of boxes (CDRSB), Alzheimer's disease assessment scale (ADAS), etc. Fusing these modalities is crucial to transforming different data into a high-quality deep representation that provides a detailed view of the patient. The literature has used these tests to improve ML models' performance to detect AD [13], [35], [36]. In our study, we used stacking models, feature selection methods, and optimization methods to provide a model that can detect different classes of AD based on heterogeneous modalities of sub-scores and heterogeneous machine learning base models. The proposed model can fuse

different outputs from different models that use different modalities to make the final prediction of classes.

The main contributions of this paper are summarized as follows:

- Integration of Heterogeneous Models and Modalities: We propose a multi-level stacking model to detect AD, including two classes (AD and cognitively normal (CN) and three classes (AD, CN, sMCI).
- Utilization of Medically Relevant Cognitive Sub-scores: We leverage the ADNI dataset and extract cognitive sub-scores that are cost-effective and medically relevant. These sub-scores provide valuable information for building robust AD detection models.
- Feature Selection Optimization: We employ swarm algorithms to optimize feature selection. By iteratively exploring the feature space, we identify the most informative subset of features from different modalities. This process enhances the model's performance and focuses on the most relevant attributes for AD detection.
- Comparative Performance Evaluation: We compare our proposed model with other classical machine learning models and stacking models using a fusion dataset. This comprehensive evaluation demonstrates that our model achieves the highest performance in AD detection among the evaluated models.
- Interpretability through SHAP: We utilize the SHAP library to provide interpretable explanations for the model's predictions. This allows us to understand the contribution of each feature towards the AD detection decision, improving the transparency and interpretability of our model.

This study is organized as follows: In Section II, the motivation for our research is briefly discussed. Section III describes our proposed method; in Section IV, the results of our experiments are discussed. Section V shows the discussion, including a comparison with related work and model explainability. In Section VI, the paper is concluded.

II. RELATED WORK

Single-modal baseline data and ML/DL models, C. Kavitha et al. [37] used different ML models: DT, RF, SVM, XGBoost, and voting classifier with different FS algorithms (correlation coefficient, information gain, and Chi-Square) on the OASIS dataset. The RF classifier and XGBoost achieved high accuracy levels. Using the same dataset, in [38], the authors compared different models, SVM, LR, DT, and RF, to predict Alzheimer's disease. The SVM model achieved high performance. In [39], the authors applied the Hybrid Feature Selection Model (CHFS) to identify the best features from the medical dataset of 1229 potential patient samples. Then, they used stacked ML models for full features and selected features. The features chosen with the stacking model achieved the highest accuracy. In [40], the authors used SVM, RF, and Gradient Boosting (GB) to predict the transition from MCI to AD using MRI and PET. The result showed that RF achieved the highest performance. In [41], the authors applied RF to predict the transition from MCI to AD using balanced data from ADNI. In [42], the authors applied the SVM DT to determine whether a patient suffers from AD or MCI. In [43], the authors used SVM to detect AD using MRI, PET, and SPECT. In [44], the authors applied RF to demographic and genetic data to predict AD.

Some research applied DL models to image datasets. For example, Islam et al. [45] a deep CNN for Alzheimer's disease diagnosis using brain MRI data analysis with four classes: CN, EMCI, LMCI, and AD. The models significantly improve multiclass classification. Rallabandi et al. [46] applied a nonlinear SVM to classify normal aging controls (MCI), early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI), and AD using whole-brain magnetic resonance imaging scans of individuals.

Using the ADNI dataset, In [47], the authors applied different ML models: NB, DT, rule induction, and generalized linear model (GLM) to determine the most distinctive aspect of AD staging using five classes CN, EMCI, LMCI, SMC, and AD. The result showed that GLM achieved the highest accuracy. Ahmed et al. [48] suggested a new model that utilized Laplacian re-decomposition for picture fusion. They divided the fused imaging into three categories, NC, MCI, and AD, by combining data from two imaging modalities, MRI and PET, along with XG-Boosting. The results showed how the approach improved competitive performance. The proposed framework outperformed NB, DT, SVM, and RF. Liu et al. [49] created a multi-model deep CNN framework for automatic hippocampal segmentation and categorization of AD. Hippocampal segmentation was done first using a deep CNN model. Then, a 3D DenseNet was developed to learn distinguishing picture features for disease categorization using the segmented hippocampal area as a starting point. The evaluation dataset was made up of T1-weighted sMRI data from the ADNI database.

Multimodal baseline data and ML/DL models: It is expected that integrating heterogeneous multimodal data (e.g., neuroimages, lab tests, memory tests, genetics, etc.) will enhance the performance of ML models and support tailored and customized decision-making [13], [50], [51], [52].

Tong et al. [53] applied ML models to different modalities: magnetic resonance imaging (MRI) and fluorodeoxyglucose (FDG) positron emission tomography (PET) to classify AD and MCI. Lodha et al. [54] proposed 3d-CNN models for an image fusion approach using MRI neuroimages to identify subjects with Alzheimer's and analyze images of brain regions connected to that disease. In [55], the goal is to increase AD and MCI detection accuracy using ensemble learning techniques incorporating three classifiers: RF, NN, and KNN. In addition, we want to increase our understanding of 11C-PIB brain regions. Weighted and unweighted ensemble methods were tested on the ADNI 11C-PIB PET imaging dataset.

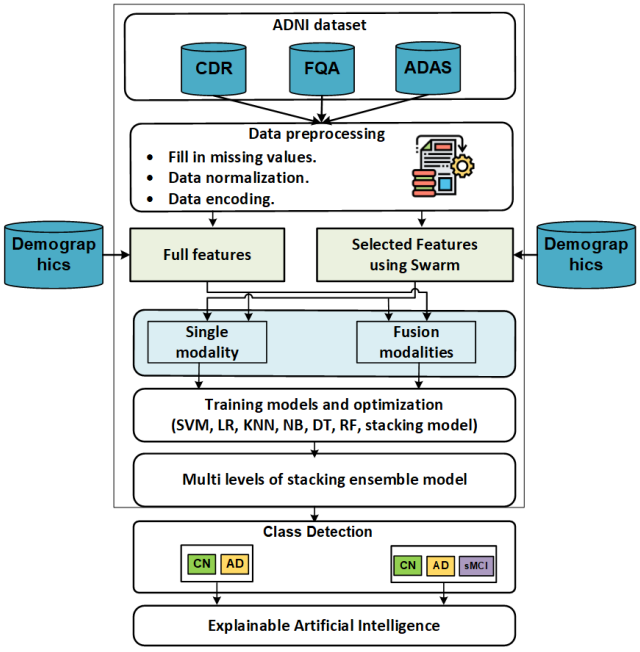


FIGURE 1. The proposed framework.

In [56], the authors investigated the role of RF in diagnosing AD with MRI images. The histogram was used to extract images' features, which were used as inputs for various classifiers, including SVM, KNN, RF, NB, LR, and DT. In [57], the authors applied DL to integrated MRI and PET image modalities to detect AD.

There are many cost-effective and cognitive (neuropsychological) scores (CSs) that a clinical expert administers to detect AD. These tests include a functional activities questionnaire (FAQ), clinical dementia rating – the sum of boxes (CDRSB), and Alzheimer's disease assessment scale (ADAS).

The previous studies did not use CSs sub-scores to detect AD; they just analyzed a summary of scores or image data. Therefore, in our paper, we developed a novel model-based three-level stacking model using sub-scores and fusion of sub-scores (multimodalities) to detect AD with two classes (AD, CN) and three classes (AD, CN, sMCI). Also, Black-box models are represented using explainable AI (XAI) for two and three classes.

III. MATERIALS AND METHOD

The paper aims to propose a multi-level stacking ensemble model to predict AD based on the sub-scores of cognitive scores. As shown in Figure 1, the proposed framework has a set of phases: data collection, data preprocessing, baseline ML models optimization, and multi-level stacking ensemble model. Each stage is described in detail as follows.

Data Collection: ADNI (Alzheimer's Disease Neuroimaging Initiative) is utilized for our study [36]. The dataset has 1363 patients, including 467 of sMCI, 418 of CN, and 478 of AD. Our study concentrates on the baseline data; 30 medical features are aggregated from three modalities, including clinical dementia rating (CDR), functional activities questionnaire test (FQR), and AD assessment scale (ADAS). In addition, the baseline dataset (i.e., AGE, PTGENDER, PTEDUCAT, PTRACCAT, PTMARRY, APOE4, FDG, ABETA, TAU, and PTAU) is fused with each modality.

A. DATA PREPROCESSING

This step aims to improve the quality of the aggregated data.

- The missing values are filled using median and mean for the numeric data and mode for the categorical data.
- Encoding data involves converting categorical features into numeric ones. Each category is encoded with a unique value. Several data encoding methods have been developed, including hot encoding, label encoding, and ordinal encoding. Our study converted categorical features to numeric features using the label encoding technique.

B. BASELINE ML MODELS OPTIMIZATION

The proposed architectural model is trained using state-of-the-art ML methods. An ensemble of these models is trained using the acquired feature data to finish the prediction process.

- Random forest (RF) is a type of Supervised Machine Learning Algorithm commonly used for classification and regression problems [58], [59]. It generates decision trees from various samples and uses their majority vote or average for classification or regression respectively [60], [61]. It can process data sets with both continuous variables, as in regression, and categorical variables, as in classification.
- Logistic regression (LR) is a statistical technique for exploring the association between a categorical dependent variable and one or more independent variables [62], [63]. In LR, the dependent variable is typically [64], whereas the independent variables can be continuous or categorical. This is accomplished by fitting the data to a logistic function, also known as a sigmoid function [65].
- Decision tree (DT) is a hierarchical decision support model that reveals options and their probable outcomes, involving chance occurrences, resource costs, and usability [66], [67]. This non-parametric, supervised learning algorithmic approach employs conditional control statements and is suitable for classification and regression applications. DT comprise root node, branches, internal nodes, and leaf nodes that form a hierarchical tree-like architecture [68].
- Support vector machine (SVM) is a supervised ML algorithm that can be utilized for both regression and classification applications [68], [69]. The main principle underlying SVM is to create a hyperplane that separates the information points into multiple classes with the most significant margin, defined as the distance between the hyperplane and the nearest data points from each class [70], [71], [72].

- Naive Bayes (NB) is a probabilistic ML algorithm founded on Bayes' theorem, which implies that the likelihood of a hypothesis as a class label given some observed evidence like the set of features proportional to the probability of the evidence given the hypothesis multiplied by the hypothesis's prior probability [73], [74].
- K-nearest neighbor (KNN) is a machine learning technique that can be used for both classification and regression purposes [75]. It is a non-parametric method, which implies that it makes no assumptions about the underlying distribution of the data. The basic principle behind KNN is to discover the K closest neighbors to the newly added information point and then use those neighbors to figure out the class or value for the newly acquired data point. The value K denotes the total number of nearest neighbors for consideration and is usually selected by the user. In the case of regression, KNN assigns the average value of the K closest neighbors [76]. When classifying data, KNN assigns the class label that appears the most frequently among the K closest neighbors [77].

C. FEATURE SELECTION OPTIMIZATION

Practical swarm optimization (PSO) is a meta-heuristic feature optimization algorithm inspired by swarms. It was developed to find the optimal solutions among solution spaces [78]. PSO simulates the behavior of birds and animals that do not have a group leader (i.e., fish schooling, bird flocking). Flocks try to reach the best solution through communication with other members in good situations. This process is iteratively repeated until the best solution is reached [79]. PSO shares several similarities with other evolutionary algorithms (i.e., Genetic algorithm); however, it differs from other optimization techniques as it has no crossover or mutation process and only depends on the objective function to find the optimal solution.

Feature selection is a critical process that significantly impacts the overall performance of a model. Our study employed the particle swarm optimization (PSO) algorithm for feature selection. The selection of PSO was based on several reasons, which can be summarized as follows. Firstly, PSO can handle both continuous and categorical variables, eliminating the need for assumptions about the underlying data distribution. This flexibility allows PSO to effectively handle datasets containing a combination of discrete and continuous variables [80]. Secondly, PSO offers the advantage of easy parallelization, enabling it to be run on multiple processors or computers simultaneously. This parallelization capability enhances the search process and speeds up the feature selection procedure, leading to more efficient and effective results [81]. Additionally, PSO is a population-based algorithm, which makes it suitable for datasets with high dimensionality and many features. By maintaining a population of candidate solutions, PSO can explore the search space more comprehensively and identify relevant features in complex, high-dimensional datasets [82], [83]. By utilizing PSO for feature selection, we address the challenges of different variable types, leverage parallelization for efficiency, and effectively handle high-dimensional datasets. This algorithm contributes to the overall performance of our model by identifying the most informative and discriminative features, thereby improving its accuracy, robustness, and generalization capabilities.

Algorithm 1 Algorithm PSO

Input: N sample side
p: problem dimension
M: Maximaxm iterations
LS: the lower bound of the search space
US: the upper bound of the search space
Output: S_{best} : the best solution

Start
Initialize the search process randomly.
for $i \leftarrow 1$ *to* N **do**
    $v_i^0 \leftarrow$ random volicty vector $[LS \cup US]^p$ // initialize the practical velocity
    $x_i^0 \leftarrow$ random position $[LS \cup US]^p$ // initialize position
    $p_{\text{best}}^0 \leftarrow x_i^0$ initialize the initialize the best solution
Apply eq (2) to get g_{best}^0
$m \leftarrow 1$
while $m \leq M$ **do**
    for $i = 1$ *to* S **do**
        $r^1 r^2 \leftarrow$ are two independent vectors that generated randomly $[0.1]^D$
        Apply eq (3) // update the velocity.
        Apply eq (4) // update the position.
        if $\mathrm{f}\left(x_i^t\right) < f\left(x_{best}^{t-1}\right)$ **then**
        $\left(x_{\text{best}\,i}^t\right) \leftarrow f\left(x_{\text{best}\,i}^t\right)$
    Apply eq 2 to get the best solution // update the overall best position.
    $m \leftarrow m + 1$

The mathematical model of PSO could be summarized in the following points: (1) each practical has a position, fitness value, and velocity. (2) each practical search for the optimal fitness value and position. (3) a list of the best position and best fitness is recorded. Algorithm 1 details the steps of PSO. PSO has various advantages over other algorithms, including (1) an efficient search algorithm. (2) it doesn't require variable scaling or standardization. (3) [84]. The algorithm of PSO is presented in 1.

D. THE PROPOSED MODEL OPTIMIZATION

Stacking ensemble learning with multimodal data refers to the blending of predictions derived from various ML models

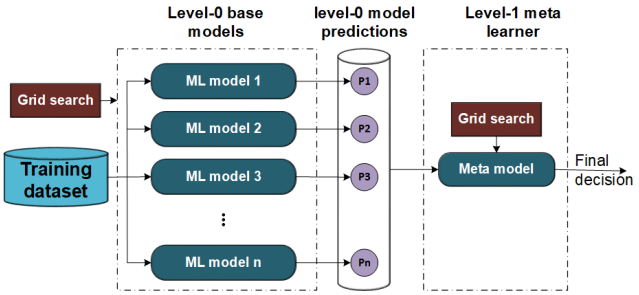


FIGURE 2. Single level stacking ensemble model.

that have been trained on different modalities or sources of data. The term ''multimodality'' refers to the utilization of various input modalities or distinct sorts of data [85], [86], [87]. This is especially beneficial when handling complex and heterogeneous data [88]. The stacking ensemble is based mainly on the dynamic weighing of several base classifiers and the learning of the best combination of their individual predictions in a way that improves the overall performance of the resulting ensemble [89]. The critical requirement for building a successful ensemble is selecting the most accurate and diverse list of base models. This combination of these models' predictions adds bias, which counters the variance of a single base model. Stacking is the most sophisticated approach for combining the predictions of base classifiers. A separate ML model called meta-learner is used to learn the predictions of base classifiers and automatically assigns weights to every base model based on its performance level [52]. Meta-learner deduces the biases of base models with respect to the training sets, so meta-learner is a weighted averaging method that assigns weights to the input predictions. This is called a single-level stacking ensemble; see Figure 2. This is the current state-of-the-art in the staking ensemble. In this paper, we propose a novel multi-level stacking ensemble where we do not select the best base model for every input modality. Still, we use all baseline models with all modalities and build N metal learners, where N is the number of modalities. This is level 1 of the process. Level 2 is to consider the resulting meta-classifiers' decisions as base classifiers. The N output decisions from Level 1 are used as input for the Level 2 meta-learner. In this case, we utilize the variance of every modality with every base model.

Our study aims to propose a multi-level stacking model as shown in Figure 3. The following is a description of each level.

- Level-based heterogeneous models with homogeneous modality:
  At level 1, each modality's initial training dataset (X) has m features. Different ML models (RF, LR, DT, SVM, KNN, and NB) are trained on X.
  At level 2, each modality model provides predictions for the outcome (y) and is stacked together in order to train and test the meta-learner.
- Level-based heterogeneous models with heterogeneous modality:

TABLE 1. Parameters of PSO.

Parameter	Value
Population size	20
Max num of generation	30
Early stopping	True
Local best weight	1
Global best weight	1
Use local random seed.	True

  At this level, the output prediction of the stacking model from each modality (ADAS, CDR, and FQA) in the training and testing datasets is merged to build a new dataset for the Level 2 stacking ensemble.
  In level 3, training stacking is used to train the level 2 meta-learner, and the testing set is used to evaluate the meta-learner and produce the final prediction.

IV. EXPERIMENTS RESULTS

This section describes the results of performing ML models and the proposed models using different sub-scores of modalities with two (AD, CN) and three classes (AD, CN, sMCI) classification problems. We named the proposed models based on the name of modality and the fusion of modalities (PMS_ADAS_CDR, PMS_CDR_FAQ, PMS_ADAS_FAQ, and PMS_ADAS_CDR_FAQ).

A. EXPERIMENT SETUP

The Scikit-learn package was used for ML. All experiments were conducted on a laptop with an Intel Core i7- 8750H CPU at 2.2 GHz and 16 G of RAM running on Windows 10 (64 bits). The dataset was divided into 80%

The values of parameters for the POS algorithm are shown in Table 1.

B. EVALUATION MODELS

All models were evaluated using different metrics, including precision (PRE), recall (REC), f1-score (F1), and accuracy (ACC). Each metric is calculated using the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) metrics. While the result was presented in (TN) as negative, it was returned as positive in (TP). As opposed to this, TP stands for positive results, and they are actually positive, while (TN) stands for negative results [90].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}. \quad (1)$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F - score = \frac{2 \cdot precision \cdot recall}{precision + recall} \quad (4)$$

C. THE RESULTS OF 2 CLASSES (AD, CN)

This section presents the results of applied models to full features and selected features by swarm with two classes (AD, CN)

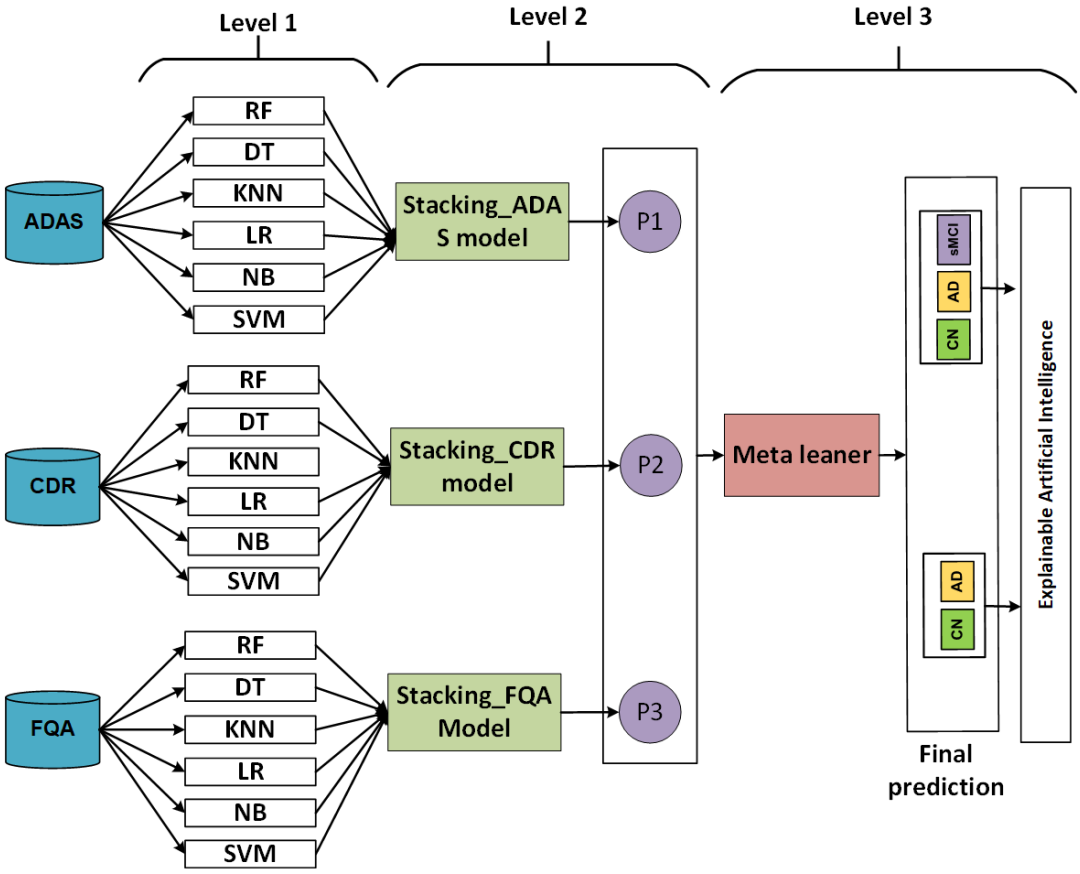


FIGURE 3. Multi-level stacking ensemble learners for predicting Alzheimer's disease.

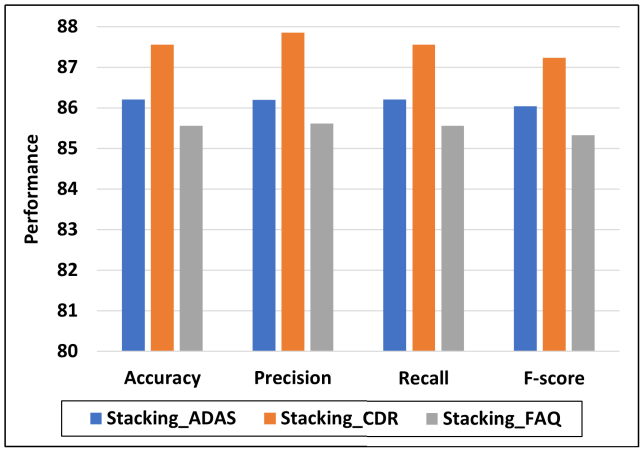
1) THE RESULTS OF FULL FEATURES

This section described the results of our proposed model against the regular ML classifiers: DT, SVM, RF, KNN, LR, NB, and stacking with two classes and full features.

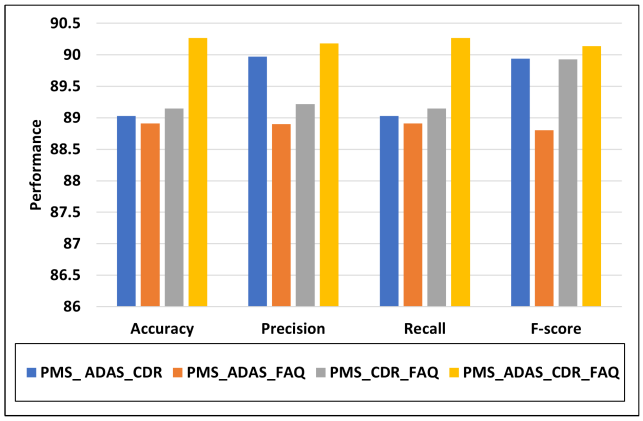
Table 2 shows the results of models using single modality and multi-modalities with two classes (AD and CN) and full features.

First, we evaluated stacking models against the regular ML classifiers using single modalities (ADAS, CDR, and FQA) with two classes and full features. We can see that, For feature set ADAS, Stacking_ADAS showed the highest performance percentage (i.e., 86.21 of ACC, 86.20 of PRE, 86.21 of REC, 86.04 of F1). DT gave the lowest performance percentage (i.e., 83.07 of ACC, 83.56 of PRE, 83.07 of REC, and 83.15 of F1). For the feature set CDR, Stacking_CDR showed the highest performance percentage (i.e., 87.56 of ACC, 87.85 of PRE, 87.56 of REC, and 87.23 of F1). DT gave the lowest rate (83.51 of ACC, 83.55 of PRE, 83.51 of REC, and 83.81 of F1). For the feature set FAQ, Stacking_FAQ showed the highest performance percentage (i.e., 85.56 of ACC, 85.61 of PRE, 85.56 of REC, and 85.33 of F1). DT gave the lowest rate (i.e., 80.70 of ACC, 80.73 of PRE, 80.70 of REC, and 80.71 of F1).

Second, we evaluated the proposed models (multi-level stacking models) against the regular ML classifiers and stacking models using multi-modalities (ADAS_CDR, ADAS_FQA, CDR_FQA, and ADAS_CDR_FQA) with two classes and full features. We can see that the performance was enhanced by 1% to 3% compared to the stacking models. For the feature set ADAS_CDR, the PMS_ADAS_CDR showed the highest performance percentage (i.e., 89.03 of ACC, 89.97 of PRE, 89.03 of REC, and 89.94 of F1). Stacking_ADAS_CDR showed the second-highest performance percentage. DT gave the lowest rate (i.e., 84.70 of ACC, 84.73 of PRE, 84.70 of REC, and 84.71 of F1). For the feature set ADAS_FAQ, the PMS_ADAS_FAQ showed the highest performance percentage (i.e., 88.91 of ACC, 88.90 of PRE, 88.91 of REC, and 88.80 of F1). Stacking_ADAS_FAQ showed the second-highest performance percentage. DT gave the lowest rate (i.e., 84.31 of ACC, 84.29 of PRE, 84.31 of REC, and 84.63 of F1). For the feature set CDR_FAQ, the PMS_CDR_FAQ showed the highest performance percentage (i.e., 89.15 of ACC, 89.22 of PRE, 89.15 of REC, and 89.93 of F1). Stacking_CDR_FAQ showed the second-highest performance percentage. NB gave the lowest rate (i.e., 85.51 of ACC, 85.55 of PRE, 85.51 of REC, and 85.81 of F1). For the feature set ADAS_CDR_FAQ, the PMS_ADAS_CDR_FAQ showed the highest performance percentage (i.e., 90.27 of ACC, 90.18 of PRE, 90.27 of REC, and 90.14 of F1). Stacking_ADAS_CDR_FAQ showed the



(a) two classes, single modality



(b) two classes, multi modalities

FIGURE 4. The best models for each modality with two classes and full features.

second-highest performance percentage. DT gave the lowest rate (i.e., 85.39 of ACC, 85.34 of PRE, 85.39 of REC, and 85.17 of F1)

Figure (a) 4 shows the best models for single modality; we noticed that features of CDR achieved the best results by Stacking_CDR. In contrast, the features of FAQ have the lowest result by Stacking_FAQ. Figure (b) 4 shows the best models for multi-modalities; we noticed that features of ADAS_CDR_FAQ achieved the best results by PMS_ADAS_CDR_FAQ. In contrast, the features of ADAS_FAQ has the lowest results.

2) THE RESULTS OF SELECTED FEATURES BY SWARM

This section described the results of our proposed model against the regular ML classifiers: DT, SVM, RF, KNN, LR, NB, and stacking with two classes and selected features by swarm.

Table 3 shows the results of models using single modality and multi-modalities with two classes (AD and CN) and selected features by swarm.

First, we evaluated stacking models against the regular ML classifiers using selected features from single modalities (ADAS, CDR, and FQA) with two classes. We can see that, For selected features set ADAS, Stacking_ADAS showed the highest performance percentage (i.e., 90.03 of ACC, 90.97 of PRE, 90.03 of REC, 90.99 of F1). DT gave the lowest performance percentage (i.e., 83.28 of ACC, 83.78 of PRE, 83.28 of REC, and 83.44 of F1). RF, LR and SVM recorded the same performance percentage. For the selected features set CDR, Stacking_CDR showed the highest performance percentage (i.e., 89.15 of ACC, 89.38 of PRE, 89.15 of REC, and 89.87 of F1). DT gave the lowest rate (81.82 of ACC, 81.89 of PRE, 81.82 of REC, and 81.85 of F1). For the selected features set FAQ, Stacking_FAQ showed the highest performance percentage (i.e., 88.98 of ACC, 88.13 of PRE, 88.98 of REC, and 88.67 of F1). DT gave the lowest rate (i.e., 82.11 of ACC, 81.96 of PRE, 82.11 of REC, and 82.02 of F1).

Second, we evaluated the proposed models (multi-level stacking models) against the regular ML classifiers and stacking models using multi-modalities (ADAS_CDR, ADAS_FQA, CDR_FQA, and ADAS_CDR_FQA) with two classes and selected features. We can see that the performance was enhanced by 1% to 3% compared to the stacking models. For the selected features set ADAS_CDR, the PMS_ADAS_CDR showed the highest performance percentage (i.e., 91.20 of ACC, 91.19 of PRE, 91.20 of REC, and 91.11 of F1). Stacking_ADAS_CDR showed the second-highest performance percentage. DT gave the lowest rate (i.e., 84.46 of ACC, 84.49 of PRE, 84.46 of REC, and 84.47 of F1). For the selected features set ADAS_FAQ, the PMS_ADAS_FAQ showed the highest performance percentage (i.e., 90.56 of ACC, 90.76 of PRE, 90.56 of REC, and 90.27 of F1). Stacking_ADAS_FAQ showed the second-highest performance percentage. DT gave the lowest rate (i.e., 85.04 of ACC, 85.22 of PRE, 85.04 of REC, and 85.11 of F1). For the selected features set CDR_FAQ, the PMS_CDR_FAQ showed the highest performance percentage (i.e., 90.85 of ACC, 90.50 of PRE, 90.85 of REC, and 90.50 of F1). Stacking_CDR_FAQ showed the second-highest performance percentage. NB gave the lowest rate (i.e., 85.92 of ACC, 85.83 of PRE, 85.92 of REC, and 85.66 of F1). For the selected features set ADAS_CDR_FAQ, the PMS_ADAS_CDR_FAQ showed the highest performance percentage (i.e., 92.08 of ACC, 92.07 of PRE, 92.08 of REC, and 92.01 of F1). Stacking_ADAS_CDR_FAQ showed the second-highest performance percentage. DT gave the lowest rate (i.e., 86.22 of ACC, 86.11 of PRE, 86.22 of REC, and 86.15 of F1).

Figure (A) 5 shows the best models for single modality; we noticed that selected features of ADAS achieved the best results by Stacking_ADAS. In contrast, the selected features of FAQ have the lowest result by Stacking_FAQ. Figure (B) 5 shows the best models for multi-modalities; we noticed that selected features of ADAS_CDR_FAQ achieved the best results by PMS_ADAS_CDR_FAQ. In contrast, the selected features of ADAS_FAQ and CDR_FAQ have the lowest results.

TABLE 2. The performance of models with the two classes and full features.

Datasets	Approaches	Models	Testing results			
			ACC	PRE	REC	F1
ADAS	Regular ML classifiers	RF	85.97	85.08	85.97	85.47
		LR	85.60	85.65	85.60	85.46
		DT	83.07	83.56	83.07	83.15
		SVM	85.60	85.59	85.60	85.47
		KNN	84.04	84.12	84.04	84.82
		NB	84.25	84.36	84.25	84.93
	Stacking model	Stacking_ADAS	86.21	86.20	86.21	86.04
CDR	Regular ML classifiers	RF	86.51	87.55	86.51	85.81
		LR	86.15	86.38	86.15	86.87
		DT	83.51	83.55	83.51	83.81
		SVM	86.68	86.07	86.68	86.27
		KNN	85.56	85.61	85.56	85.33
		NB	85.27	85.40	85.27	85.98
	Stacking model	Stacking_CDR	87.56	87.85	87.56	87.23
FAQ	Regular ML classifiers	RF	83.87	84.45	83.87	83.07
		LR	82.10	82.16	82.10	82.78
		DT	80.70	80.73	80.70	80.71
		SVM	83.80	83.97	83.80	83.42
		KNN	83.34	83.21	83.34	83.25
		NB	82.63	82.64	82.63	82.26
	Stacking model	Stacking_FAQ	85.56	85.61	85.56	85.33
ADAS_CDR	Regular ML classifiers	RF	86.87	86.45	86.87	86.07
		LR	85.15	85.09	85.15	85.02
		DT	84.70	84.73	84.70	84.71
		SVM	86.86	86.80	86.86	86.71
		KNN	85.74	85.78	85.74	85.56
		NB	85.68	85.62	85.68	85.49
	Stacking models	Stacking_ADAS_CDR	88.15	88.17	88.15	88.96
	Muti-level Stacking model	PMS_ ADAS_CDR	89.03	89.97	89.03	89.94
ADAS_FAQ	Regular ML classifiers	RF	86.58	86.70	86.58	86.96
		LR	85.26	85.34	85.26	85.11
		DT	84.31	84.29	84.31	84.63
		SVM	86.46	86.61	86.46	86.24
		KNN	85.28	85.29	85.28	85.08
		NB	85.97	85.67	85.97	85.51
	Stacking models	Stacking_ADAS_FAQ	87.56	87.85	87.56	87.23
	Muti-level Stacking model	PMS_ADAS_FAQ	88.91	88.90	88.91	88.80
CDR_FAQ	Regular ML classifiers	RF	87.39	87.95	87.39	86.90
		LR	89.15	89.22	89.15	88.93
		DT	85.51	85.55	85.51	85.81
		SVM	88.27	88.40	88.27	87.98
		KNN	87.10	87.01	87.10	87.89
		NB	86.92	86.83	86.92	86.66
	Stacking models	Stacking_CDR_FAQ	88.56	88.68	88.56	88.30
	Muti-level Stacking model	PMS_CDR_FAQ	89.15	89.22	89.15	89.93
ADAS_CDR_FAQ	Regular ML classifiers	RF	88.44	88.65	88.44	88.19
		LR	88.44	88.41	88.44	88.30
		DT	85.39	85.34	85.39	85.17
		SVM	88.56	88.61	88.56	88.33
		KNN	87.10	86.98	87.10	86.99
		NB	86.80	86.70	86.80	86.61
	Stacking model	Stacking_ADAS_CDR_FAQ	89.86	89.78	89.86	89.79
	Muti-level Stacking model	PMS_ADAS_CDR_FAQ	90.27	90.18	90.27	90.14

D. THE RESULTS OF 3 CLASSES (AD, CN, SMCI)

This section presents the results of applied models to full features and selected features by swarm with three classes (AD, CN, sMCI)

1) THE RESULTS OF FULL FEATURES

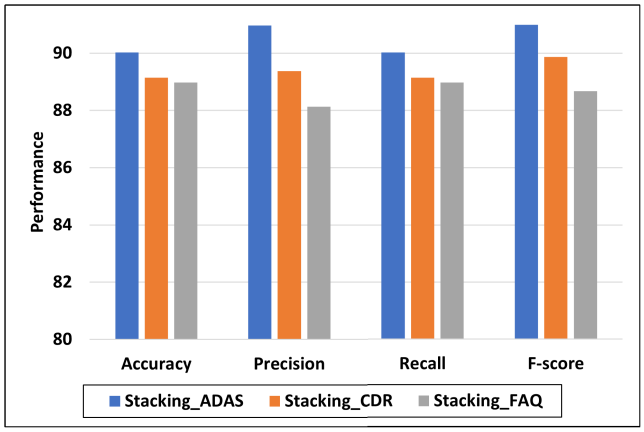
We evaluated our proposed model against the regular ML classifiers: DT, SVM, RF, KNN, LR, NB, XGB, and stacking. Table 4 shows the results of models using single modality and

TABLE 3. The performance of models with the two classes and selected features by swarm.

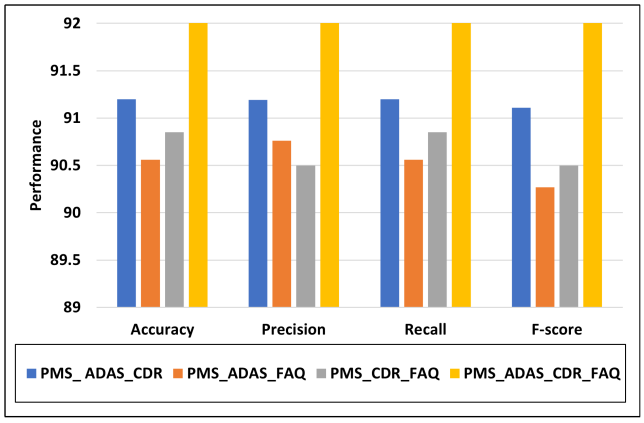
Datasets	Approaches	Models	Testing results			
			ACC	PRE	REC	F1
ADAS	Regular ML classifiers	RF	85.97	85.08	85.97	85.47
		LR	85.60	85.65	85.60	85.46
		DT	83.07	83.56	83.07	83.15
		SVM	85.60	85.59	85.60	85.47
		KNN	84.04	84.12	84.04	84.82
		NB	84.25	84.36	84.25	84.93
	Stacking model	Stacking_ADAS	86.21	86.20	86.21	86.04
CDR	Regular ML classifiers	RF	86.51	87.55	86.51	85.81
		LR	86.15	86.38	86.15	86.87
		DT	83.51	83.55	83.51	83.81
		SVM	86.68	86.07	86.68	86.27
		KNN	85.56	85.61	85.56	85.33
		NB	85.27	85.40	85.27	85.98
	Stacking model	Stacking_CDR	87.56	87.85	87.56	87.23
FAQ	Regular ML classifiers	RF	83.87	84.45	83.87	83.07
		LR	82.10	82.16	82.10	82.78
		DT	80.70	80.73	80.70	80.71
		SVM	83.80	83.97	83.80	83.42
		KNN	83.34	83.21	83.34	83.25
		NB	82.63	82.64	82.63	82.26
	Stacking model	Stacking_FAQ	85.56	85.61	85.56	85.33
ADAS_CDR	Regular ML classifiers	RF	86.87	86.45	86.87	86.07
		LR	85.15	85.09	85.15	85.02
		DT	84.70	84.73	84.70	84.71
		SVM	86.86	86.80	86.86	86.71
		KNN	85.74	85.78	85.74	85.56
		NB	85.68	85.62	85.68	85.49
	Stacking models	Stacking_ADAS_CDR	88.15	88.17	88.15	88.96
	Muti-level Stacking model	PMS_ ADAS_CDR	89.03	89.97	89.03	89.94
ADAS_FAQ	Regular ML classifiers	RF	86.58	86.70	86.58	86.96
		LR	85.26	85.34	85.26	85.11
		DT	84.31	84.29	84.31	84.63
		SVM	86.46	86.61	86.46	86.24
		KNN	85.28	85.29	85.28	85.08
		NB	85.97	85.67	85.97	85.51
	Stacking models	Stacking_ADAS_FAQ	87.56	87.85	87.56	87.23
	Muti-level Stacking model	PMS_ADAS_FAQ	88.91	88.90	88.91	88.80
CDR_FAQ	Regular ML classifiers	RF	87.39	87.95	87.39	86.90
		LR	89.15	89.22	89.15	88.93
		DT	85.51	85.55	85.51	85.81
		SVM	88.27	88.40	88.27	87.98
		KNN	87.10	87.01	87.10	87.89
		NB	86.92	86.83	86.92	86.66
	Stacking models	Stacking_CDR_FAQ	88.56	88.68	88.56	88.30
	Muti-level Stacking model	PMS_CDR_FAQ	89.15	89.22	89.15	89.93
ADAS_CDR_FAQ	Regular ML classifiers	RF	88.44	88.65	88.44	88.19
		LR	88.44	88.41	88.44	88.30
		DT	85.39	85.34	85.39	85.17
		SVM	88.56	88.61	88.56	88.33
		KNN	87.10	86.98	87.10	86.99
		NB	86.80	86.70	86.80	86.61
	Stacking models	Stacking_ADAS_CDR_FAQ	89.86	89.78	89.86	89.79
	Muti-level Stacking model	PMS_ADAS_CDR_FAQ	90.27	90.18	90.27	90.14

multi-modalities with three classes (AD, CN, and sMCI) and full features.

First, we evaluated stacking models against the regular ML classifiers using single modalities (ADAS, CDR, and FQA) with three classes and full features. We can see that, For feature set ADAS, Stacking_ADAS showed the highest performance percentage (i.e., 71.55 of ACC, 71.27 of PRE, 71.55 of REC, 71.34 of F1). KNN gave the lowest

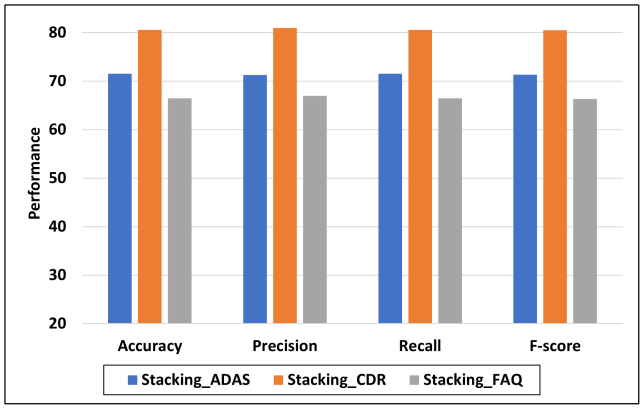


(a) two classes, single modality

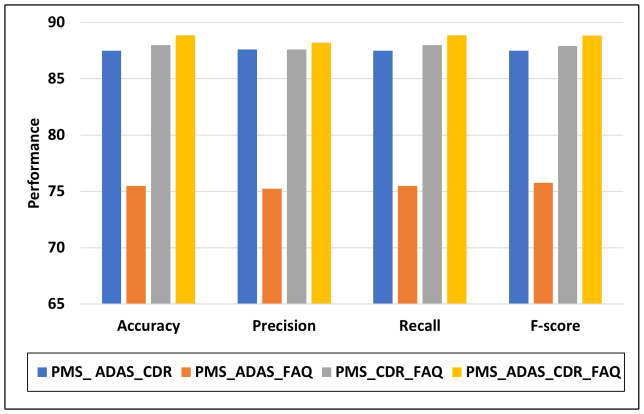


(b) two classes, multi modalities

FIGURE 5. The best models for each modality with two classes and selected features.



(a) two classes, single modality



(b) two classes, multi modalities

FIGURE 6. The best models for each modality with three classes and full features.

performance percentage (i.e., 62.76 of ACC, 62.29 of PRE, 62.76 of REC, and 62.29 of F1). For the feature set CDR, Stacking_CDR showed the highest performance percentage (i.e., 84.56 of ACC, 84.96 of PRE, 84.56 of REC, and 84.52 of F1). KNN gave the lowest rate (76.83 of ACC, 76.44 of PRE, 76.83 of REC, and 76.52 of F1). For the feature set FAQ, Stacking_FAQ showed the highest performance percentage (i.e., 66.50 of ACC, 66.99 of PRE, 66.50 of REC, and 66.37 of F1). KNN gave the lowest rate (i.e., 61.58 of ACC, 60.98 of PRE, 61.58 of REC, and 60.99 of F1).

Second, we evaluated the proposed models (multi-level stacking models) against the regular ML classifiers and stacking models using multi-modalities (ADAS_CDR, ADAS_FQA, CDR_FQA, and ADAS_CDR_FQA) with two classes and full features. We can see that the performance was enhanced by 1% to 3% compared to the stacking models. For the feature set ADAS_CDR, the PMS_ADAS_CDR showed the highest performance percentage (i.e., 87.50 of ACC, 87.59 of PRE, 87.50 of REC, and 87.49 of F1). Stacking_ADAS_CDR showed the second-highest performance percentage. NB gave the lowest rate (i.e., 72.14 of ACC, 72.79 of PRE, 72.14 of REC, and 72.75 of F1). For the feature set ADAS_FAQ, the PMS_ADAS_CDR showed the highest performance percentage (i.e., 75.49 of ACC, 75.25 of PRE, 75.49 of REC, and 75.77 of F1). Stacking_ADAS_FAQ showed the second-highest performance percentage. KNN gave the lowest rate (i.e., 63.64 of ACC, 63.72 of PRE, 63.64 of REC, and 63.30 of F1). For the feature set CDR_FAQ, the PMS_CDR_FAQ showed the highest performance percentage (i.e., 87.98 of ACC, 87.59 of PRE, 87.98 of REC, and 87.91 of F1). Stacking_CDR_FAQ showed the second-highest performance percentage. KNN gave the lowest rate (i.e., 65.69 of ACC, 65.21 of PRE, 65.69 of REC, and 65.61 of F1). For the feature set ADAS_CDR_FAQ, the PMS_ADAS_CDR_FAQ showed the highest performance percentage (i.e., 88.86 of ACC, 88.20 of PRE, 88.86 of REC, and 88.82 of F1). Stacking_ADAS_CDR_FAQ showed the second-highest performance percentage. NB gave the lowest rate (i.e., 70.04 of ACC, 70.96 of PRE, 70.04 of REC, and 70.49 of F1).

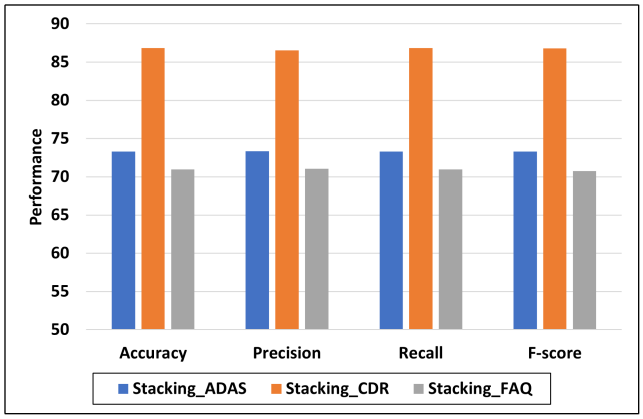
Figure (A) 6 shows the best models for single modality; we noticed that features of CDR achieved the best results by Stacking_CDR. In contrast, the features of FAQ have the lowest result by Stacking_FAQ. Figure (B) 6 shows the best models for multi-modalities; we noticed that features of ADAS_CDR_FAQ achieved the best results by PMS_ADAS_CDR_FAQ. In contrast, the features of ADAS_FAQ have the lowest results by PMS_ADAS_FAQ.

TABLE 4. The performance of models with the three classes and full features.

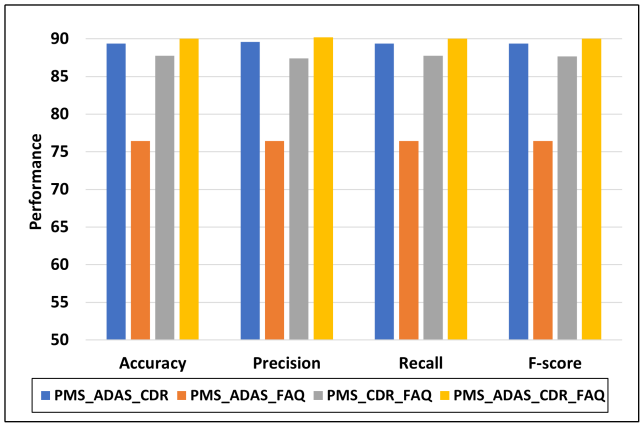
Datasets	Approaches	Models	Testing results			
			ACC	PRE	REC	F1
ADAS	Regular ML classifiers	RF	70.90	70.55	70.90	70.65
		LR	69.09	69.93	69.09	69.98
		DT	67.18	67.71	67.18	67.25
		SVM	68.97	68.58	68.97	86.67
		KNN	62.76	62.29	62.76	62.29
		NB	65.98	68.36	65.98	65.66
	Stacking model	Stacking_ADAS	71.55	71.27	71.55	71.34
CDR	Regular ML classifiers	RF	79.11	79.68	79.11	79.69
		LR	79.56	79.96	79.56	79.52
		DT	77.69	77.39	77.69	77.19
		SVM	78.68	78.84	78.68	78.54
		KNN	76.83	76.44	76.83	76.52
		NB	77.13	79.30	77.13	77.89
	Stacking model	Stacking_CDR	80.56	80.96	80.56	80.52
FAQ	Regular ML classifiers	RF	64.22	65.20	64.22	64.92
		LR	64.21	64.74	64.21	64.18
		DT	63.30	63.38	63.30	63.55
		SVM	65.45	65.74	65.45	65.26
		KNN	61.58	60.98	61.58	60.99
		NB	62.46	62.47	62.46	62.48
	Stacking model	Stacking_FAQ	66.50	66.99	66.50	66.37
ADAS_CDR	Regular ML classifiers	RF	85.20	85.58	85.20	85.18
		LR	85.03	85.15	85.03	85.02
		DT	83.98	83.03	83.98	83.00
		SVM	85.91	85.34	85.91	85.88
		KNN	75.37	75.30	75.37	75.14
		NB	72.14	72.79	72.14	72.75
	Stacking models	Stacking_ADAS_CDR	86.20	86.48	86.20	86.18
	Muti-level Stacking model	PMS_ ADAS_CDR	87.50	87.59	87.50	87.49
ADAS_FAQ	Regular ML classifiers	RF	72.49	72.49	72.49	72.83
		LR	72.61	72.99	72.61	72.69
		DT	70.67	71.03	70.67	70.46
		SVM	72.31	72.45	72.31	72.64
		KNN	68.91	69.63	68.91	68.70
		NB	63.64	63.72	63.64	63.30
	Stacking models	Stacking_ADAS_FAQ	74.19	74.40	74.19	74.28
	Muti-level Stacking model	PMS_ADAS_FAQ	75.49	75.25	75.49	75.77
CDR_FAQ	Regular ML classifiers	RF	84.86	84.72	84.86	84.77
		LR	84.98	84.67	84.98	84.99
		DT	83.98	83.80	83.98	83.97
		SVM	84.27	84.83	84.27	84.21
		KNN	75.07	75.02	75.07	74.70
		NB	65.69	65.21	65.69	65.61
	Stacking models	Stacking_CDR_FAQ	85.15	85.67	85.15	85.10
	Muti-level Stacking model	PMS_CDR_FAQ	87.98	87.59	87.98	87.91
ADAS_CDR_FAQ	Regular ML classifiers	RF	86.86	86.56	86.86	86.78
		LR	86.39	86.28	86.39	86.46
		DT	85.39	85.63	85.39	85.24
		SVM	86.44	86.80	86.44	86.41
		KNN	80.94	81.69	80.94	81.04
		NB	70.04	70.96	70.04	70.49
	Stacking models	Stacking_ADAS_CDR_FAQ	87.44	87.70	87.44	87.42
	Muti-level Stacking model	PMS_ADAS_CDR_FAQ	88.86	88.20	88.86	88.82

2) THE RESULTS OF SELECTED FEATURES BY SWARM

This section described the results of our proposed model against the regular ML classifiers: DT, SVM, RF, KNN, LR, NB, and stacking with two classes and selected features by swarm. Table 5 shows the results of models using single modality and multi-modalities with three classes (AD,



(a) two classes, single modality



(b) two classes, multi modalities

FIGURE 7. The best models for each modality with three classes and selected features.

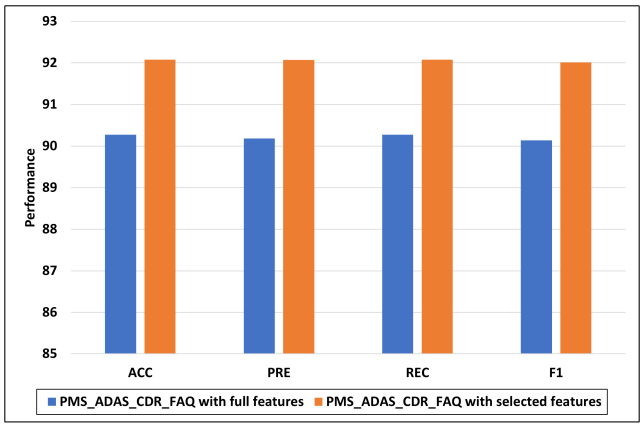


FIGURE 8. The best-proposed models for two classes with full features and selected features.

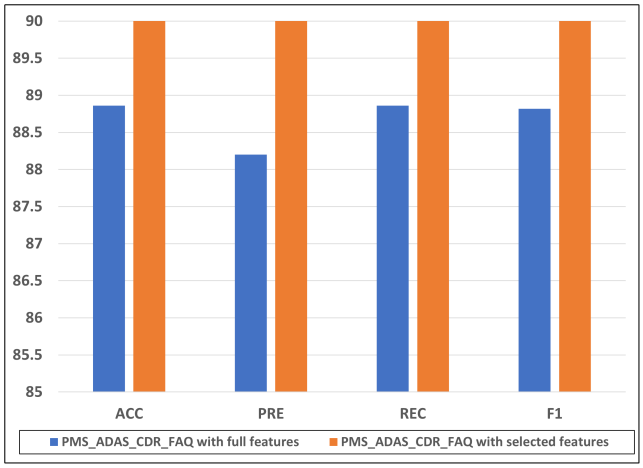


FIGURE 9. Comparing between best model with full features and selected features.

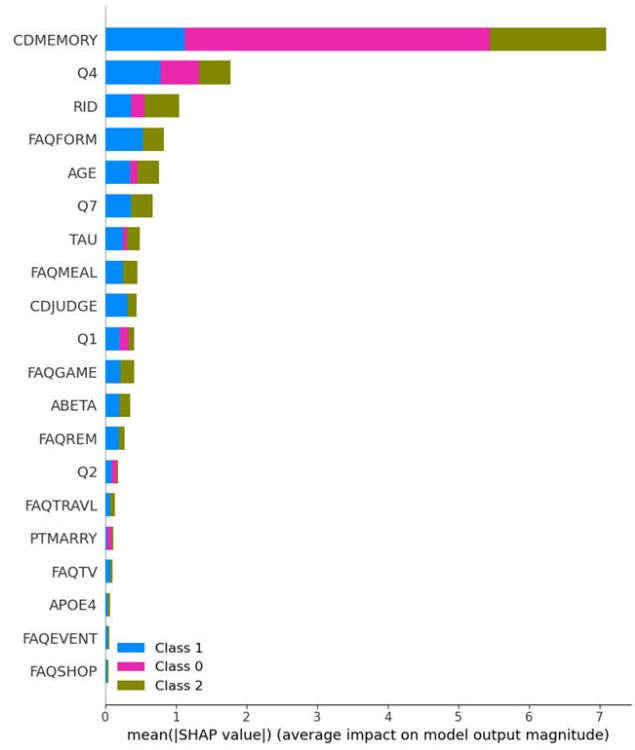


FIGURE 10. Summary plot for three class problems (AD=0, CN=1, sMCI=2).

CN, and SMCI) and selected features by swarm. First, we evaluated stacking models against the regular ML classifiers using selected features from single modalities (ADAS, CDR, and FQA) with three classes. For the selected features set ADAS, Stacking_ADAS showed the highest performance percentage (73.31 of ACC, 73.36 of PRE, 73.31 of REC, 73.32 of F1). NB gave the lowest performance percentage (66.57 of ACC, 68.85 of PRE, 66.57 of REC, and 66.27 of F1). For the selected features set CDR, Stacking_CDR showed the highest performance percentage (86.86 of ACC, 86.56 of PRE, 86.86 of REC, and 86.78 of F1). DT gave the lowest rate (81.82 of ACC, 81.83 of PRE, 81.82 of REC, and 81.82 of F1). For the selected features set FAQ, Stacking_FAQ showed the highest performance percentage (70.97 of ACC, 71.06 of PRE, 70.97 of REC, and 70.77 of F1). DT gave the lowest rate (64.52 of ACC, 64.19 of PRE, 64.52 of REC, and 64.07 of F1). Second,

TABLE 5. The performance of models with the three classes and selected features.

Datasets	Approaches	Models	Testing results			
			Accuracy	Precision	Recall	F1-score
ADAS	Regular ML classifiers	RF	71.02	71.70	71.02	71.84
		LR	70.09	69.93	70.09	69.98
		DT	68.33	68.11	68.33	68.62
		SVM	70.97	70.58	70.97	70.67
		KNN	67.16	66.68	67.16	65.71
		NB	66.57	68.85	66.57	66.27
	Stacking model	Stacking_ADAS	73.31	73.36	73.31	73.32
CDR	Regular ML classifiers	RF	85.74	85.04	85.74	85.71
		LR	85.56	85.96	85.56	85.52
		DT	81.82	81.83	81.82	81.81
		SVM	85.68	85.84	85.68	85.54
		KNN	78.30	78.48	78.30	77.93
		NB	82.68	82.93	82.68	82.60
	Stacking model	Stacking_CDR	86.86	86.56	86.86	86.78
FAQ	Regular ML classifiers	RF	68.67	68.17	68.67	68.87
		LR	69.21	69.74	69.21	69.18
		DT	64.52	64.19	64.52	64.07
		SVM	67.45	68.74	67.45	67.26
		KNN	65.98	67.74	65.98	65.02
		NB	65.10	67.43	65.10	63.79
	Stacking model	Stacking_FAQ	70.97	71.06	70.97	70.77
ADAS_CDR	Regular ML classifiers	RF	87.91	87.14	87.91	87.89
		LR	87.03	87.15	87.03	87.02
		DT	83.34	83.46	83.34	83.39
		SVM	87.91	87.34	87.91	87.88
		KNN	78.89	79.51	78.89	78.32
		NB	84.80	84.95	84.80	84.81
	Stacking model	Stacking_ADAS_CDR	88.91	88.96	88.91	88.91
	Muti-level Stacking model	PMS_ADAS_CDR	89.38	89.58	89.38	89.36
ADAS_FAQ	Regular ML classifiers	RF	72.61	72.44	72.61	72.93
		LR	72.61	72.99	72.61	72.69
		DT	70.97	71.00	70.97	70.94
		SVM	72.31	72.45	72.31	72.64
		KNN	66.86	68.65	66.86	66.26
		NB	69.79	69.72	69.79	69.90
	Stacking model	Stacking_ADAS_FAQ	74.78	74.38	74.78	75.01
	Muti-level Stacking model	PMS_ADAS_FAQ	76.45	76.43	76.45	76.43
CDR_FAQ	Regular ML classifiers	RF	85.50	85.40	85.50	85.40
		LR	85.98	85.67	85.98	85.99
		DT	84.51	84.68	84.51	84.57
		SVM	85.27	85.83	85.27	85.21
		KNN	76.25	76.76	76.25	76.04
		NB	83.92	83.69	83.92	83.04
	Stacking model	Stacking_CDR_FAQ	86.56	86.07	86.56	86.51
	Muti-level Stacking model	PMS_CDR_FAQ	87.74	87.40	87.74	87.68
ADAS_CDR_FAQ	Regular ML classifiers	RF	88.74	88.96	88.74	88.72
		LR	88.15	88.65	88.15	88.21
		DT	84.63	84.00	84.63	84.67
		SVM	87.56	87.52	87.56	87.51
		KNN	76.54	76.71	76.54	76.43
		NB	84.92	85.51	85.92	85.06
	Stacking model	Stacking_ADAS_CDR_FAQ	89.44	89.90	89.44	89.40
	Muti-level Stacking model	PMS_ADAS_CDR_FAQ	90.03	90.19	90.03	90.05

we evaluated the proposed models (multi-level stacking models) against the regular ML classifiers and stacking models using multi-modalities (ADAS_CDR, ADAS_FQA, CDR_FQA, and ADAS_CDR_FQA) with two classes and selected features. We can see that the performance was enhanced by 1% to 3% compared to the stacking models. For the selected features set ADAS_CDR, the PMS_ADAS_CDR model showed the highest performance percentage (89.38 of ACC, 89.58 of PRE, 89.38 of REC, and 89.36 of F1). Stacking_ADAS_CDR showed the second-highest performance percentage. DT gave the lowest rate (83.34 of ACC, 83.46 of PRE, 83.34 of REC, and 83.39 of F1). For the selected features set ADAS_FAQ, the PMS_ADAS_FAQ model achieved the highest performance percentage (76.45 of ACC, 76.43 of PRE, 76.45 of REC, and 76.43 of F1). Stacking_ADAS_FAQ showed the second-highest performance percentage. KNN gave the lowest rate (66.86 of ACC, 68.65 of PRE, 66.86 of REC, and 66.26 of F1). For the

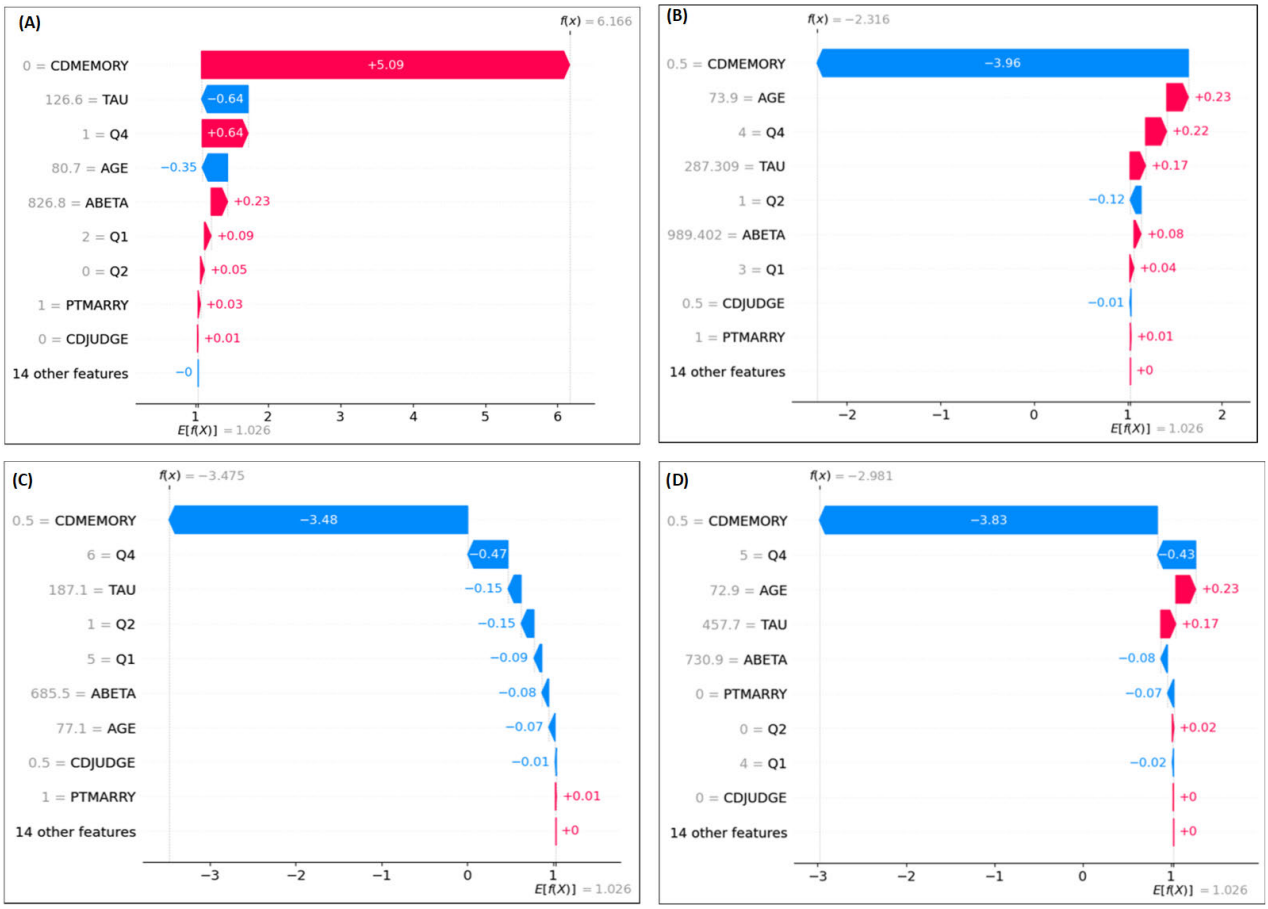


FIGURE 11. Waterfall for three classes problem.

selected features set CDR_FAQ, the PMS_CDR_FAQ model achieved the highest performance percentage (89.74 of ACC, 87.40 of PRE, 87.74 of REC, and 87.68 of F1). Stacking_CDR_FAQ showed the second-highest performance percentage. KNN gave the lowest rate (76.25 of ACC, 76.76 of PRE, 76.25 of REC, and 76.04 of F1). For the selected features set ADAS_CDR_FAQ, the PMS_ADAS_CDR_FAQ model achieved the highest performance percentage (90.03 of ACC, 90.19 of PRE, 90.03 of REC, and 90.05 of F1). Stacking_ADAS_CDR_FAQ showed the second-highest performance percentage. KNN gave the lowest rate (76.54 of ACC, 76.71 of PRE, 76.54 of REC, and 76.43 of F1).

Figure (A) 7 shows the best models for single modality; we noticed that features of ADAS achieved the best results by Stacking_CDR. In contrast, the features of FAQ have the lowest result by Stacking_FAQ. Figure (B) 7 shows the best models for multi-modalities; we noticed that features of ADAS_CDR_FAQ achieved the best results by PMS_ADAS_CDR_FAQ. In contrast, the features of ADAS_FAQ have the lowest results by PMS_ADAS_FAQ.

V. DISCUSSION

This section shows the discussion, including comparing related work and model explainability.

A. THE BEST MODELS OF DETECTING AD WITH DIFFERENT CLASSES TASK

From the results in the subsection, the fusion of different modalities achieved the highest performance with different classes. Also, the proposed model has achieved the highest performance.

In Figure 8, we can see that PMS_ADAS_CDR_FAQ with two classes has the highest performance with selected features compared to full features.

In Figure 9, we can see that PMS_ADAS_CDR_FAQ with three classes has the highest performance with selected features compared to full features.

B. EXPLAINABLE ARTIFICIAL INTELLIGENCE

To ensure the robustness of the developed model and make it fully trusted from a medical expert's perspective, we proved the XAI capabilities to understand why the model made certain decisions and which features had the most significant impact on these decisions. We utilized SHAP and LIME explainers to interpret the proposed classifier. In this section, we use the SHAP library to explain the developed decisions on both the global and instance levels.

1) THREE CLASSES (AD, CN, sMCI)

In this section, we concentrate on the explanation of the three-class problem. First, we utilized the summary plot to show the

TABLE 6. comparing with previous studies that used different datasets to detect AD.

Research study	Dataset	Models	Classes	The performance
[37]	OASIS	XBoost	non-demented demented	ACC=85.12, PRE=83, REC=83 and F1=85
[38]	OASIS	SVM	non-demented demented	ACC=92 and REC=91.89
[39]	OASIS dataset	CHFS+SVM	non-demented demented	ACC=96.50%, REC=96.5
[45]	ADNI	Deep CNN	CN, EMCI, LMCI and AD	ACC=93
[46]	ADNI	SVM	CN, EMCI, LMCI and AD	REC=75 F1=72
[47]	ADNI	GLM	CN, EMCI, LMCI, SMC and AD	ACC=88.24
[48]	MRI and PET	XGB	NC, MCI, and AD	ACC=98.06%
[39]	Multi-modality (MRI,FDG, and PET)	ML models	AD and CN	ACC= 94.8%
Our work	Sub-scores of fusion (FQA, CDR, and ADAS)	Muti-level of stacking models	AD and CN	ACC=92.08 PRE=92.07 REC=92.08 F1=92.01

impact of all features in predicting the three classes. Figure 10 shows the CD Memory, Q7, FAQFORM, and CD_jUDGE Form have essential features. Blue, purple, and green colors represent the three classes. CD_memory is significant in predicting where Q4 is critical in predicting AD and less essential for sMCI class. Figure 11 (A, B, C, D) shows the importance of each feature according to each instance. A physician can easily measure if the developed model makes an accurate decision and which parts are considered to make the final decision. As shown in Figure 11 (A, B, C, D), each plot shows the Base_value, which represents the value according to the whole dataset, and the predict_proba_value that the probability according to the specific instance, the left side shows the feature values and arrows show the feature contribution. The final probability for each instance is calculated by adding the base value to the feature contributions.

2) TWO CLASSES (AD AND CN)

First, we concentrate on the two-class problem; we create a summary plot that identifies the role of each feature in the overall model decisions. The x-axis shows the critical features and the y-axis shows the feature importance in a bar graph; each bar has a length equal to its importance. The blue color clarifies how the model contributes toward the AD class, whereas the red color shows how the model contributes to the CN class. According to Figure 12, CDMemory, Q7, and Q4 forms are considered the most important features. To explore the importance of each feature according to instance level, we used SHAP explainers via a waterfall plot. The waterfall plot shows all features contributing to the developed decision sorted according to SHAP values. As shown in Figure 13 (A, B, C, D), each plot shows the Base_value, which represents the value according to the whole dataset, and the predict_proba_value that the probability according to

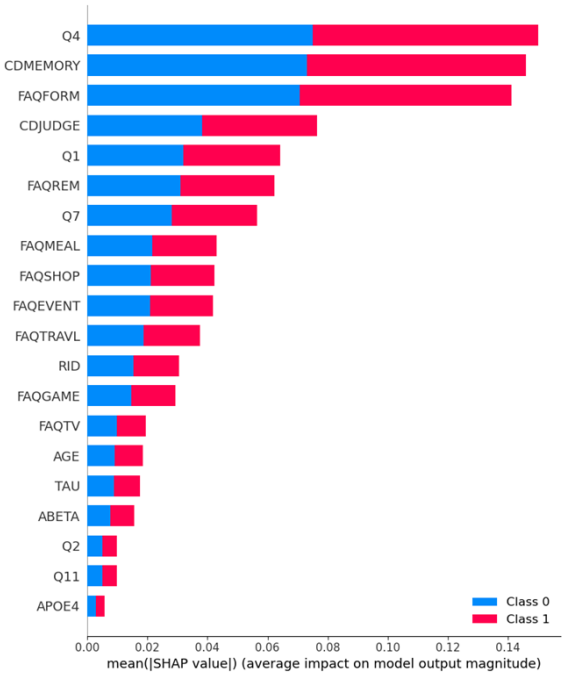


FIGURE 12. Summary plot for two class problems (0=AD, 1=CN).

the specific instance, the left side shows the feature values according to that instance and the arrows show the feature contribution according to prediction. Each row is according to negative and positive contributions in blue and red bars. Each row shows how each feature affects negatively or positively. These explanations can help medical experts understand and trust the model's decisions.

To uphold the accuracy of the XAI from a scientific point of view, we employ a meticulous analysis based on a medical lens. This analysis ascertains that the explanations generated by our model align with the existing understanding within the realm of medicine.

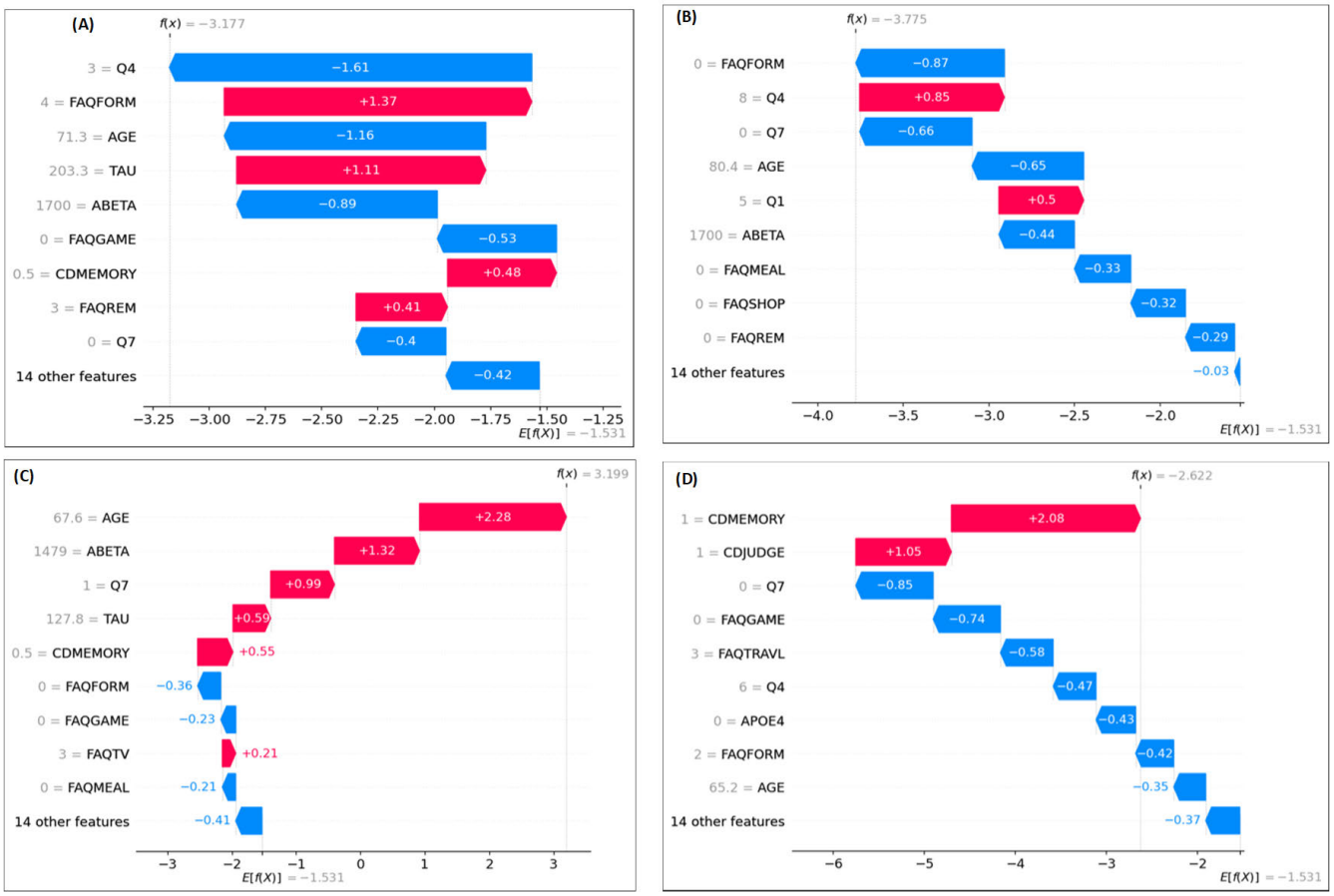


FIGURE 13. Waterfall for two classes problem.

As we can see in Figures 13 and 12 for two class Q4 and CD_memory is the most significant features. These features are clinically confirmed in several studies. For example, in [91], authors assured that CD_memory is the earliest and most prominent symptom experienced by individuals with AD. The same is true in [92], which also ensures that analyzing memory performance plays a crucial role in defining different subtypes of AD, assessing disease progression, and making predictions about its prognosis. ADAS score incorporates biomarkers derived from cerebrospinal fluid analysis or measurements of hippocampal volume [93]. And we can see in Figures 10 and 11 for three class that Q4, age, CD_memory, and Q3 has a significant impact on the overall decision; that result is also confirmed in [94], they explore the intricate relationship between aging and Alzheimer's disease, they examine and consolidate evidence at various levels, including molecular, cellular, and systemic.

C. COMPARING WITH PREVIOUS STUDIES THAT USED DIFFERENT DATASETS OF AD TO DETECT AD

Table 6 compares previous studies that used different AD datasets to detect AD. Some of the authors used OASIS with non-demented and demented classes. For example, [37] XBoost was recorded (ACC=85.12, PRE=83, RE=83 and F1=85). In [38], SVM recorded 92 ACC and 91.89 REC. In [39], CHFS+SVM was registered ACC=96.50, REC=96.5. The authors used the ADNI dataset with four classes: CN, EMCI, LMCI, and AD. In [45], Deep CNN was recorded ACC=93, In [46], SVM was recorded REC=75 and F1=72,. In [47], GLM was recorded ACC=88.24. Other studies used muti-modalities to detect AD [39], [48]. All studies have not used sub-scores or multi-level stacking models based on heterogeneous models with heterogeneous modalities and XAI.

VI. CONCLUSION

We have proposed a new multi-level stacking model based on the sub-scores of CSs and the fusion of these sub-scores from the ADNI dataset to predict AD with higher accuracy. The problem has been formulated at three different levels of complexity: two classes (AD, CN) and three classes (AD, CN, sMCI). The novel multi-level stacking includes Level-based heterogeneous models with homogeneous modality and Level-based heterogeneous models with heterogeneous modality. Firstly, each modality's initial training dataset (X) has m features. Different ML models (RF, LR, DT, SVM, KNN, and NB) are trained on X. Then, each modality model provides predictions for the outcome (y) and is stacked to train and test the meta-learner. Secondly, the output prediction of the stacking model from each

modality (ADAS, CDR, and FQA) in the training and testing datasets is merged to build a new dataset for the Level 2 stacking ensemble. Training stacking is employed to train the level 2 meta-learner, and the testing set is utilized to evaluate the meta-learner and produce the final prediction. Practical swarm optimization (PSO) selects the best features from each sub-score. The proposed model achieved the best results compared to single-level stacking and classical ML models using different datasets and fused datasets with full features and selected features. The results showed that the multi-modalities recorded the best performance compared to the single modality. In addition, the proposed models (multi-level stacking models) achieved the highest performance with selected features against the regular ML classifiers and stacking models using full multi-modalities with two classes and three classes (accuracy=92.08, precision=92.07, recall=92.08, and F1-score=92.01) and (accuracy=90.03, precision=90.19, recall=90.03, and F1-score=90.05), respectively.

DATA AVAILABILITY

All datasets used to support the findings of this study are available from the direct link in the dataset citations.

CONFLICT OF INTERESTS

All authors declare that they have no conflicts of interest.

REFERENCES

[1] A. Association, "2018 Alzheimer's disease facts and figures," *Alzheimer's Dementia*, vol. 14, no. 3, pp. 367–429, Mar. 2018.

[2] S. Al-Shoukry, T. H. Rassem, and N. M. Makbol, "Alzheimer's diseases detection by using deep learning algorithms: A mini-review," *IEEE Access*, vol. 8, pp. 77131–77141, 2020.

[3] R. H. Blank and R. H. Blank, "Alzheimer's disease and other dementias: An introduction," in *Social & Public Policy of Alzheimer's Disease in the United States*, 2019, pp. 1–26.

[4] *World Health Origination*. Accessed: 2023. [Online]. Available: https://www.who.int/news/item/07-12-2017-dementia-number-of-people-affected-to-triple-in-next-30-years

[5] J. Neugroschl and S. Wang, "Alzheimer's disease: Diagnosis and treatment across the spectrum of disease severity," *Mount Sinai J. Med., J. Transl. Personalized Med.*, vol. 78, no. 4, pp. 596–612, Jul. 2011.

[6] M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, and R. C. Petersen, "The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease," *Focus*, vol. 11, no. 1, pp. 96–106, 2013.

[7] G. M. McKhann, "The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's Dementia*, vol. 7, no. 3, pp. 263–269, 2011.

[8] R. A. Sperling, "Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's Dementia*, vol. 7, no. 3, pp. 280–292, 2011.

[9] M. A. Lovell, "A potential role for alterations of zinc and zinc transport proteins in the progression of Alzheimer's disease," *J. Alzheimer's Disease*, vol. 16, no. 3, pp. 471–483, Mar. 2009.

[10] D. Siedlecki-Wullich, J. Catalá-Solsona, C. Fábregas, I. Hernández, J. Clarimon, A. Lleó, M. Boada, C. A. Saura, J. Rodríguez-Álvarez, and A. J. Miñano-Molina, "Altered microRNAs related to synaptic function as potential plasma biomarkers for Alzheimer's disease," *Alzheimer's Res. Therapy*, vol. 11, pp. 1–11, Dec. 2019.

[11] M. Tanveer, B. Richhariya, R. U. Khan, A. H. Rashid, P. Khanna, M. Prasad, and T. C. Lin, "Machine learning techniques for the diagnosis of Alzheimer's disease: A review," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 16, no. 1s, pp. 1–35, Apr. 2020.

[12] K. D. Tzimourta, V. Christou, A. T. Tzallas, N. Giannakeas, L. G. Astrakas, P. Angelidis, D. Tsalikakis, and M. G. Tsipouras, "Machine learning algorithms and statistical approaches for Alzheimer's disease analysis based on resting-state EEG recordings: A systematic review," *Int. J. Neural Syst.*, vol. 31, no. 5, May 2021, Art. no. 2130002.

[13] S. El-Sappagh, H. Saleh, R. Sahal, T. Abuhmed, S. M. R. Islam, F. Ali, and E. Amer, "Alzheimer's disease progression detection model based on an early fusion of cost-effective multimodal data," *Future Gener. Comput. Syst.*, vol. 115, pp. 680–699, Feb. 2021.

[14] X. Hong, R. Lin, C. Yang, N. Zeng, C. Cai, J. Gou, and J. Yang, "Predicting Alzheimer's disease using LSTM," *IEEE Access*, vol. 7, pp. 80893–80901, 2019.

[15] A. S. Alatrany, A. J. Hussain, J. Mustafina, and D. Al-Jumeily, "Machine learning approaches and applications in genome wide association study for Alzheimer's disease: A systematic review," *IEEE Access*, vol. 10, pp. 62831–62847, 2022.

[16] D. Aarsland, K. Brønnick, J. Larsen, O. Tysnes, and G. Alves, "Cognitive impairment in incident, untreated Parkinson disease: The Norwegian Parkwest study," *Neurology*, vol. 72, no. 13, pp. 1121–1126, 2009.

[17] M. Perovnik, P. Tomše, J. Jamšek, A. Emeršič, C. Tang, D. Eidelberg, and M. Trošt, "Identification and validation of Alzheimer's disease-related metabolic brain pattern in biomarker confirmed Alzheimer's dementia patients," *Sci. Rep.*, vol. 12, no. 1, Jul. 2022, Art. no. 11752.

[18] J. H. Park, H. E. Cho, J. H. Kim, M. M. Wall, Y. Stern, H. Lim, S. Yoo, H. S. Kim, and J. Cha, "Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data," *NPJ Digit. Med.*, vol. 3, no. 1, p. 46, 2020.

[19] O. Sagi and L. Rokach, "Ensemble learning: A survey," *WIREs Data Mining Knowl. Discovery*, vol. 8, no. 4, Jul. 2018, Art. no. e1249.

[20] A. Almulihi, H. Saleh, A. M. Hussien, S. Mostafa, S. El-Sappagh, K. Alnowaiser, A. A. Ali, and M. Refaat Hassan, "Ensemble learning based on hybrid deep learning model for heart disease early prediction," *Diagnostics*, vol. 12, no. 12, p. 3215, Dec. 2022.

[21] S. Shafieian and M. Zulkernine, "Multi-layer stacking ensemble learners for low footprint network intrusion detection," *Complex Intell. Syst.*, vol. 9, pp. 3787–3799, Aug. 2023.

[22] Y. Fan, N. Batmanghelich, C. M. Clark, and C. Davatzikos, "Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline," *NeuroImage*, vol. 39, no. 4, pp. 1731–1743, Feb. 2008.

[23] T. Abuhmed, S. El-Sappagh, and J. M. Alonso, "Robust hybrid deep learning models for Alzheimer's progression detection," *Knowl.-Based Syst.*, vol. 213, Feb. 2021, Art. no. 106688.

[24] O. V. Forlenza, M. Radanovic, L. L. Talib, I. Aprahamian, B. S. Diniz, H. Zetterberg, and W. F. Gattaz, "Cerebrospinal fluid biomarkers in Alzheimer's disease: Diagnostic accuracy and prediction of dementia," *Alzheimer's Dementia: Diagnosis, Assessment Disease Monitor.*, vol. 1, no. 4, pp. 455–463, Dec. 2015.

[25] T. Grimmer, C. Wutz, P. Alexopoulos, A. Drzezga, S. Förster, H. Förstl, O. Goldhardt, M. Ortner, C. Sorg, and A. Kurz, "Visual versus fully automated analyses of 18F-FDG and amyloid PET for prediction of dementia due to Alzheimer disease in mild cognitive impairment," *J. Nucl. Med.*, vol. 57, no. 2, pp. 204–207, Feb. 2016.

[26] S. El-Sappagh, J. M. Alonso-Moral, T. Abuhmed, F. Ali, and A. Bugarín-Diz, "Trustworthy artificial intelligence in Alzheimer's disease: State of the art, opportunities, and challenges," *Artif. Intell. Rev.*, vol. 56, pp. 11149–11296, Mar. 2023.

[27] Y. Huang, J. Xu, Y. Zhou, T. Tong, and X. Zhuang, "Diagnosis of Alzheimer's disease via multi-modality 3D convolutional neural network," *Frontiers Neurosci.*, vol. 13, p. 509, May 2019.

[28] P. Kumar, S. Malik, and B. Raman, "Interpretable multimodal emotion recognition using hybrid fusion of speech and image data," *Multimedia Tools Appl.*, pp. 1–22, Sep. 2023.

[29] A. A. Jihad and A. S. Abdalkafor, "A framework for sentiment analysis in Arabic text," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 16, no. 3, pp. 1482–1489, 2019.

[30] H. Saleh, S. Mostafa, L. A. Gabralla, A. O. Aseeri, and S. El-Sappagh, "Enhanced Arabic sentiment analysis using a novel stacking ensemble of hybrid and deep learning models," *Appl. Sci.*, vol. 12, no. 18, p. 8967, Sep. 2022.

[31] Y. Lyu, X. Yu, L. Zhang, and D. Zhu, "Classification of mild cognitive impairment by fusing neuroimaging and gene expression data: Classification of mild cognitive impairment by fusing neuroimaging and gene expression data," in *Proc. 14th Pervasive Technol. Rel. Assistive Environ. Conf.*, 2021, pp. 26–32.

[32] N. Rahim, S. El-Sappagh, S. Ali, K. Muhammad, J. Del Ser, and T. Abuhmed, "Prediction of Alzheimer's progression based on multimodal deep-learning-based fusion and visual explainability of time-series data," *Inf. Fusion*, vol. 92, pp. 363–388, Apr. 2023.

[33] N. El-Rashidy, S. El-Sappagh, T. Abuhmed, S. Abdelrazek, and H. M. El-Bakry, "Intensive care unit mortality prediction: An improved patient-specific stacking ensemble model," *IEEE Access*, vol. 8, pp. 133541–133564, 2020.

[34] S. Ali, S. El-Sappagh, F. Ali, M. Imran, and T. Abuhmed, "Multitask deep learning for cost-effective prediction of Patient's length of stay and readmission state using multimodal physical activity sensory data," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 12, pp. 5793–5804, Dec. 2022.

[35] S. Iddi, D. Li, P. S. Aisen, M. S. Rafii, W. K. Thompson, M. C. Donohue, and A. D. N. Initiative, "Predicting the course of Alzheimer's progression," *Brain Informat.*, vol. 6, pp. 1–18, Dec. 2019.

[36] J. C. Morris, S. Weintraub, H. C. Chui, J. Cummings, C. DeCarli, S. Ferris, N. L. Foster, D. Galasko, N. Graff-Radford, E. R. Peskind, D. Beekly, E. M. Ramos, and W. A. Kukull, "The uniform data set (UDS): Clinical and cognitive variables and descriptive data from Alzheimer disease centers," *Alzheimer Disease Associated Disorders*, vol. 20, no. 4, pp. 210–216, 2006.

[37] C. Kavitha, V. Mani, S. R. Srividhya, O. I. Khalaf, and C. A. T. Romero, "Early-stage Alzheimer's disease prediction using machine learning models," *Frontiers Public Health*, vol. 10, p. 240, Mar. 2022.

[38] M. B. Antor, A. Jamil, M. Mamtaz, M. M. Khan, S. Aljahdali, M. Kaur, P. Singh, and M. Masud, "A comparative analysis of machine learning algorithms to predict Alzheimer's disease," *J. Healthcare Eng.*, vol. 2021, Jul. 2021, Art. no. 9917919.

[39] A. A. Farid, G. I. Selim, and H. A. A. Khater, "Applying artificial intelligence techniques to improve clinical diagnosis of Alzheimer's disease," *Eur. J. Eng. Sci. Technol.*, vol. 3, no. 2, pp. 58–79, 2020.

[40] J. F. Beltrán, B. M. Wahba, N. Hose, D. Shasha, and R. P. Kline, "Inexpensive, non-invasive biomarkers predict Alzheimer transition using machine learning analysis of the Alzheimer's disease neuroimaging (ADNI) database," *PLoS ONE*, vol. 15, no. 7, Jul. 2020, Art. no. e0235663.

[41] M. Velazquez, Y. Lee, and A. D. N. Initiative, "Random forest model for feature-based Alzheimer's disease conversion prediction from early mild cognitive impairment subjects," *PLoS ONE*, vol. 16, no. 4, 2021, Art. no. e0244773.

[42] A. B. Rabeh, F. Benzarti, and H. Amiri, "Diagnosis of Alzheimer diseases in early step using SVM (support vector machine)," in *Proc. 13th Int. Conf. Comput. Graph., Imag. Visualizat. (CGiV)*, Mar. 2016, pp. 364–367.

[43] P. J. Moore, T. J. Lyons, and J. Gallacher, "Random forest prediction of Alzheimer's disease using pairwise selection from time series data," *PLoS ONE*, vol. 14, no. 2, Feb. 2019, Art. no. e0211558.

[44] H. Wang, Y. Shen, S. Wang, T. Xiao, L. Deng, X. Wang, and X. Zhao, "Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease," *Neurocomputing*, vol. 333, pp. 145–156, Mar. 2019.

[45] J. Islam and Y. Zhang, "Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks," *Brain Informat.*, vol. 5, no. 2, pp. 1–14, Dec. 2018.

[46] V. P. S. Rallabandi, K. Tulpule, and M. Gattu, "Automatic classification of cognitively normal, mild cognitive impairment and Alzheimer's disease using structural MRI analysis," *Informat. Med. Unlocked*, vol. 18, Jan. 2020, Art. no. 100305.

[47] M. Shahbaz, S. Ali, A. Guergachi, A. Niazi, and A. Umer, "Classification of Alzheimer's disease using machine learning techniques," in *Proc. Data*, 2019, pp. 296–303.

[48] H. Ahmed, H. Soliman, S. El-Sappagh, T. Abuhmed, and M. Elmogy, "Early detection of Alzheimer's disease based on Laplacian redecomposition and XGBoosting," *Comput. Syst. Sci. Eng.*, vol. 46, no. 3, pp. 2773–2795, 2023.

[49] M. Liu, F. Li, H. Yan, K. Wang, Y. Ma, L. Shen, and M. Xu, "A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease," *NeuroImage*, vol. 208, Mar. 2020, Art. no. 116459.

[50] S. El-Sappagh, J. M. Alonso, S. M. R. Islam, A. M. Sultan, and K. S. Kwak, "A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease," *Sci. Rep.*, vol. 11, no. 1, pp. 1–26, Jan. 2021.

[51] B. Cheng, M. Liu, D. Zhang, B. C. Munsell, and D. Shen, "Domain transfer learning for MCI conversion prediction," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 7, pp. 1805–1817, Jul. 2015.

[52] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 55, no. 3, pp. 856–867, Apr. 2011.

[53] L. Xu, X. Wu, K. Chen, and L. Yao, "Multi-modality sparse representation-based classification for Alzheimer's disease and mild cognitive impairment," *Comput. Methods Programs Biomed.*, vol. 122, no. 2, pp. 182–190, Nov. 2015.

[54] P. Lodha, A. Talele, and K. Degaonkar, "Diagnosis of Alzheimer's disease using machine learning," in *Proc. 4th Int. Conf. Comput. Commun. Control Autom. (ICCUBEA)*, Aug. 2018, pp. 1–4.

[55] S. Lahmiri, "Integrating convolutional neural networks, kNN, and Bayesian optimization for efficient diagnosis of Alzheimer's disease in magnetic resonance images," *Biomed. Signal Process. Control*, vol. 80, Feb. 2023, Art. no. 104375.

[56] E. Alickovic and A. Subasi, "Automatic detection of Alzheimer disease based on histogram and random forest," in *Proc. Int. Conf. Med. Biol. Eng.* Banja Luka, Bosnia Herzegovina: Springer, 2020, pp. 91–96.

[57] A. Ortiz, J. Munilla, J. M. Górriz, and J. Ramírez, "Ensembles of deep learning architectures for the early diagnosis of the Alzheimer's disease," *Int. J. Neural Syst.*, vol. 26, no. 07, Nov. 2016, Art. no. 1650025.

[58] Y. Liu, Y. Wang, and J. Zhang, "New machine learning algorithm: Random forest," in *Proc. Int. Conf. Inf. Comput. Appl.* Chengde, China: Springer, Sep. 2012, pp. 246–252.

[59] M. R. Segal, "Machine learning benchmarks and random forest regression," Tech. Rep., 2004.

[60] C. Zhang and Y. Ma, *Ensemble Machine Learning: Methods and Applications*. Cham, Switzerland: Springer, 2012.

[61] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[62] G. Tripepi, K. Jager, F. Dekker, and C. Zoccali, "Linear and logistic regression analysis," *Kidney Int.*, vol. 73, no. 7, pp. 806–810, 2008.

[63] J. J. Guido, P. C. Winters, and A. B. Rains, "Logistic regression basics," M.S. thesis, Univ. Rochester Med. Center, Rochester, NY, USA, 2006.

[64] J. E. King, "Binary logistic regression," in *Best Practices in Quantitative Methods*, 2008, pp. 358–384.

[65] H.-A. Park, "An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain," *J. Korean Acad. Nursing*, vol. 43, no. 2, pp. 154–164, 2013.

[66] A. Navada, A. N. Ansari, S. Patil, and B. A. Sonkamble, "Overview of use of decision tree algorithms in machine learning," in *Proc. IEEE Control Syst. Graduate Res. Colloq.*, Jun. 2011, pp. 37–42.

[67] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 1, pp. 20–28, Mar. 2021.

[68] H. H. Patel and P. Prajapati, "Study and analysis of decision tree based classification algorithms," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 10, pp. 74–78, Oct. 2018.

[69] M. Somvanshi, P. Chavan, S. Tambade, and S. V. Shinde, "A review of machine learning techniques using decision tree and support vector machine," in *Proc. Int. Conf. Comput. Commun. Control Autom. (ICCUBEA)*, Aug. 2016, pp. 1–7.

[70] A. Pradhan, "Support vector machine—A survey," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 2, no. 8, pp. 82–85, 2012.

[71] L. Wang, *Support Vector Machines: Theory and Applications*, vol. 177. Cham, Switzerland: Springer, 2005.

[72] A. Garba, S. Wu, and S. Khalid, "Federated search techniques: An overview of the trends and state of the art," *Knowl. Inf. Syst.*, vol. 65, pp. 5065–5095, Jul. 2023.

[73] G. I. Webb, E. Keogh, and R. Miikkulainen, "Naïve Bayes," in *Encyclopedia of Machine Learning and Data Mining*, vol. 15, 2010, pp. 713–714.

[74] E. Frank, L. Trigg, G. Holmes, and I. H. Witten, "Naive Bayes for regression," *Mach. Learn.*, vol. 41, pp. 5–25, Oct. 2000.

[75] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.
[76] P. Cunningham and S. J. Delany, "K-nearest neighbour classifiers—A tutorial," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–25, Jul. 2022.
[77] M. Steinbach and P.-N. Tan, "kNN: K-nearest neighbors," in *The Top Ten Algorithms Data Mining*. London, U.K.: Chapman & Hall, 2009, pp. 165–176.
[78] A. G. Gad, "Particle swarm optimization algorithm and its applications: A systematic review," *Arch. Comput. Methods Eng.*, vol. 29, no. 5, pp. 2531–2561, Aug. 2022.
[79] D. P. Rini, S. M. Shamsuddin, and S. S. Yuhaniz, "Particle swarm optimization: Technique, system and challenges," *Int. J. Comput. Appl.*, vol. 14, no. 1, pp. 19–26, 2011.
[80] H. Garg, "A hybrid PSO-GA algorithm for constrained optimization problems," *Appl. Math. Comput.*, vol. 274, pp. 292–305, Feb. 2016.
[81] K. Premalatha and A. Natarajan, "Hybrid pso and ga for global maximization," *Int. J. Open Problems Compt. Math*, vol. 2, no. 4, pp. 597–608, 2009.
[82] X.-S. Yang, "Genetic algorithms," in *Nature-Inspired Optimization Algorithms*, 2014, pp. 77–87.
[83] X.-S. Yang, "Particle swarm optimization," in *Nature-Inspired Optimization Algorithms*, vol. 7, 2014, pp. 99–110.
[84] S. B. Sakri, N. B. A. Rashid, and Z. M. Zain, "Particle swarm optimization feature selection for breast cancer recurrence prediction," *IEEE Access*, vol. 6, pp. 29637–29647, 2018.
[85] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 689–696.
[86] R. J. Perrin, A. M. Fagan, and D. M. Holtzman, "Multimodal techniques for diagnosis and prognosis of Alzheimer's disease," *Nature*, vol. 461, no. 7266, pp. 916–922, Oct. 2009.
[87] M. Dyrba, M. Grothe, T. Kirste, and S. J. Teipel, "Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM," *Hum. Brain Mapping*, vol. 36, no. 6, pp. 2118–2131, Jun. 2015.
[88] M. Dyrba, F. Barkhof, A. Fellgiebel, M. Filippi, L. Hausner, K. Hauenstein, T. Kirste, and S. J. Teipel, "Predicting prodromal Alzheimer's disease in subjects with mild cognitive impairment using machine learning classification of multimodal multicenter diffusion-tensor and magnetic resonance imaging data," *J. Neuroimag.*, vol. 25, no. 5, pp. 738–747, Sep. 2015.
[89] M. Lorenzi, I. J. Simpson, A. F. Mendelson, S. B. Vos, M. J. Cardoso, M. Modat, J. M. Schott, and S. Ourselin, "Multimodal image analysis in Alzheimer's disease via statistical modelling of non-local intensity correlations," *Sci. Rep.*, vol. 6, no. 1, pp. 1–8, Apr. 2016.
[90] S. Khalid, S. Wu, and F. Zhang, "A multi-objective approach to determining the usefulness of papers in academic search," *Data Technol. Appl.*, vol. 55, no. 5, pp. 734–748, Oct. 2021.
[91] H. Jahn, "Memory loss in Alzheimer's disease," in *Dialogues in Clinical Neuroscience*, 2022.
[92] A. Porsteinsson, R. Isaacson, S. Knox, M. Sabbagh, and I. Rubino, "Diagnosis of early Alzheimer's disease: Clinical practice in 2021," *J. Prevention Alzheimer's Disease*, vol. 8, pp. 371–386, Jun. 2021.
[93] R. Yaari, A. S. Fleisher, and P. N. Tariot, "Updates to diagnostic guidelines for Alzheimer's disease," *Primary Care Companion CNS Disorders*, vol. 13, no. 5, 2011, Art. no. 26971.
[94] X. Xia, Q. Jiang, J. McDermott, and J. J. Han, "Aging and Alzheimer's disease: Comparison and associations from molecular to system level," *Aging Cell*, vol. 17, no. 5, Oct. 2018, Art. no. e12802.

ABDULAZIZ ALMOHIMEED received the master's degree from Monash University, Australia, and the Ph.D. degree from the University of Southampton, U.K. He is currently an Assistant Professor with the College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia. His research interests include natural language processing, artificial intelligence, data science, the Internet of Things, and network security. He is passionate about leveraging technology to create innovative solutions.

REDHWAN M. A. SAAD received the Ph.D. degree in internet infrastructure security from University Sains Malaysia (USM). Currently he works at the College of Informatics, Midocean University, Comoros. He was a Senior Lecturer at Ibb University, Yemen. In addition, he is a Postdoctoral Research Fellow at Computer Engineering Department, Cairo University. His current research interests include cybersecurity, the Internet of Things security, intrusion detection system (IDS), intrusion prevention system (IPS), and IPv6 security.

SHERIF MOSTAFA received the master's degree in securing wireless networks from the Computer Science Department, Faculty of Computers and Information, Helwan University, and the Ph.D. degree in computer science and in medical image compression from the Computer Science Department, Faculty of Science, South Valley University, Hurghada, in 2020. He worked for four years with the Modern Sciences and Arts University (MSA) and ten years with the Department of Computer Engineering, Al Baha College, Saudi Arabia. He is currently an Assistant Professor with the Faculty of Computers and Artificial Intelligence, South Valley University. His research interests include image processing and the use of deep learning methods and computer vision.

NORA MAHMOUD EL-RASHIDY received the M.S. and Ph.D. degrees from the Faculty of Computer Science and Information, Mansoura University, Egypt, in 2020 and 2016, respectively. She is currently an Associate Professor with the Machine Learning and Information Retrieval Department, Faculty of Artificial Intelligence, Kafrelsheikh University. She is also an Assistant Professor with Galala University, an Assistant Professor with the Machine Learning and Information Retrieval Department, Kafrelsheikh University, and the Manager of the Measurement and Assessment Central Unit, Kafrelsheikh University. Her research interests include machine learning, medical informatics, distributed and hybrid clinical decision support systems, big data, and cloud computing. She is a reviewer in many journals, and she is very interested in disease diagnosis and treatment research.

SARAH FARRAG received the master's and Ph.D. degrees from the Computer Science Department, Faculty of Science, South Valley University, Egypt, in 2016 and 2021, respectively. She is currently an Assistant Professor with the Faculty of Computers and Information, South Valley University. Her research interests include artificial intelligence, machine learning, deep learning, big data analytics, data mining, sentiment analysis, and streaming data.

ABDELKAREEM GABALLAH is currently pursuing the degree with Kafrelkhaskieh University. He brings together academic excellence and practical experience with Repotech, a renowned tech company. His expertise spans deep learning in natural language processing (NLP) and computer vision, with a focus on LLMs and Arabic NLP. Beyond tech, his passion extends to generative art, showcasing his creative approach to algorithms. His multidisciplinary journey positions him as a rising talent poised to shape the future of AI and technology.

MOHAMED ABD ELAZIZ received the B.S. and M.S. degrees in computer science and the Ph.D. degree in mathematics and computer science from Zagazig University, Egypt, in 2008, 2011, and 2014, respectively. Currently he works as an Assistant Professor at the Faculty of Computer Science and Engineering, Galala University, Suez, Egypt. From 2008 to 2011, he was an Assistant Lecturer at the Department of Computer Science. Since 2014, he has been a Lecturer with the Mathematical Department, Zagazig University. He is the author of more than 100 articles. His research interests include machine learning, signal processing, image processing, and metaheuristic techniques.(Based on document published on October 2022).

SHAKER EL-SAPPAGH received the bachelor's and master's degrees in computer science from the Department of Information Systems, Faculty of Computers and Information, Cairo University, Egypt, in 1997 and 2007, respectively, and the Ph.D. degree in computer science from the Department of Information Systems, Faculty of Computers and Information, Mansoura University, Mansoura, Egypt, in 2015. In 2003, he joined as a Teaching Assistant with the Department of Information Systems, Faculty of Computers and Information, Minia University, Egypt. Since June 2016, he has been an Assistant Professor with the Department of Information Systems, Faculty of Computers and Information, Benha University. From 2018 to 2020, he was a Research Professor with the UWB Wireless Communications Research Center, Department of Information and Communication Engineering, Inha University, South Korea. He was a Research Professor with the Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain, in 2021. He is currently an Associate Professor with Galala University, Egypt. He is also a Senior Researcher with the College of Computing and Informatics, Sungkyunkwan University, South Korea. His publications in clinical decision support systems and semantic intelligence. His current research interests include machine learning, medical informatics, (fuzzy) ontology engineering, distributed and hybrid clinical decision support systems, semantic data modeling, fuzzy expert systems, and cloud computing. He is a reviewer of many journals and he is very interested in the diseases' diagnosis and treatment research.

HAGER SALEH received the Ph.D. degree from the Faculty of Computers and Information, Minia University, Egypt. She is currently an Assistant Professor with the Faculty of Computers and Artificial Intelligence, South Valley University, Hurghada, Egypt. She was a Senior Machine Learning Engineer with the Data Science for Smart Healthcare Project, from 2019 to 2023. In 2018, she was a data scientist in an analytics patrol company in Egypt. She also worked on different research projects and wrote the complete code for each one. Her research interests include machine learning, medical informatics, distributed and hybrid clinical decision support systems, big data, and cloud computing. She is a reviewer in many journals, and she is very interested in disease diagnosis and treatment research.