IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 May 2024, accepted 10 May 2024, date of publication 15 May 2024, date of current version 22 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3401104

RESEARCH ARTICLE

ML-Powered Handwriting Analysis for Early Detection of Alzheimer's Disease

UDDALAK MITRA[1] **AND SHAFIQ UL REHMAN**[2]
[1]Siliguri Institute of Technology, Siliguri, West Bengal 734009, India
[2]College of Information Technology, Kingdom University, Riffa 3903, Bahrain

Corresponding author: Shafiq Ul Rehman (s.rehman@ku.edu.bh)

This work was supported in part by Kingdom University, Bahrain, under Grant 2024-1-001.

ABSTRACT Alzheimer's disease (AD) is a progressive, incurable condition leading to decline of nerve cells and cognitive functions over time. Early detection is essential for improving quality of life, as treatment strategies aim to decelerate its progression. AD also impacts fine motor control, including handwriting. Utilizing machine learning (ML) with efficient data analysis methods for early detection of Alzheimer's disease (AD) through handwriting analysis holds considerable promise for clinical diagnosis, albeit a challenging undertaking. In this study, we address this complexity by employing ensemble machine learning, which amalgamates diverse ML algorithms to enhance predictive performance. Our approach involves developing an ensemble model for analysis of handwriting kinetics, utilizing the stacking technique to integrate multiple base-level classifiers. The study encompasses 174 individuals, including 89 diagnosed with Alzheimer's disease and 85 healthy participants, drawn from the DARWIN dataset (Diagnosis AlzheimeR WIth haNdwriting). To discern the most effective base classifiers, we employ both Repeated-k-fold and Monte-Carlo Cross Validation techniques. Subsequently, top k features are selected for each best-performing base classifier using analysis of variance (ANOVA) and recursive feature elimination (RFE). The final step involves consolidating predictions from the base classifiers through a stacking ensemble, resulting in an impressive performance. The ensemble model achieves 97.14% accuracy, 95% sensitivity, 100% specificity, 100% precision, 97.44% F1-score, 94.37% Matthews Correlation Coefficient (MCC), 94.21% Cohen Kappa, and 97.5% Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Comparative performance analysis demonstrates that our proposed model surpasses all state-of-the-art models based on the DARWIN dataset for Alzheimer's disease prediction. These findings underscore the potential of machine learning to offer highly accurate predictions for Alzheimer's disease in an affordable and non-invasive manner, emphasizing its significant clinical utility, particularly through handwriting analysis.

INDEX TERMS Alzheimer's disease prediction, diagnosis AlzheimeR WIth handwriting, ensemble machine learning, handwriting analysis for identifying neurodegenerative disease, machine learning for disease prediction, machine learning based Alzheimer's disease prediction.

I. INTRODUCTION

Dementia, a global health concern affecting over 55 million individuals worldwide. It is particularly prevalent in low- and middle-income countries, encompassing more than 60% of affected individuals, with nearly 10 million new cases reported annually [1]. Being the seventh leading cause of mortality, dementia poses a significant burden on global health, contributing substantially to disability and dependency among the elderly population. In 2019 alone, the economic impact of dementia exceeded 1.3 trillion US dollars, with half attributed to informal caregivers, who invest an average of 5 hours per day, underscoring the gravity of this public health challenge. Notably, women bear a disproportionate burden, experiencing higher disability-adjusted life years and mortality from dementia, while also providing 70% of care hours for individuals with dementia.

Neurodegenerative disorders like Alzheimer's disease (AD), account for a substantial proportion of dementia cases, comprising 60–70% of instances [1]. AD initiates a gradual

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

decline in cognitive abilities, commencing with episodic memory impairment often associated with dysfunction in the ventromedial temporal lobe [2]. As the disease advances, it leads to progressive amnesia and broader cognitive deterioration, reflecting widespread neural damage. Unfortunately, curative treatments for AD are currently lacking, and existing therapies primarily aim to impede its progression. Given the increasing global lifespan, the prevalence of AD is anticipated to rise significantly, emphasizing the critical need for enhanced clinical methods for its early detection.

In light of the intricate link between cognitive and motor functions in planning and executing movements [3], the analysis of handwriting, requiring precise motor control, emerges as an affordable and non-invasive means of evaluating neurodegenerative disease progression [4], [5], [6], [7], [8]. Leveraging machine learning techniques to assess motor function has shown promise in streamlining clinical evaluations [9], [10], [11], [12]. Handwriting tests, conducted using commonplace graphic tablets, enable the collection of kinematic and dynamic data related to handwriting and drawing activities [8]. As a result, researchers are increasingly exploring machine learning-driven approaches to recognize handwriting [13], [14], [15], [16] as well as analyzing handwriting for the diagnosis of neurodegenerative disorders, with methodologies proposed for both Alzheimer's disease [17] and Parkinson's disease [18].

A. BACKGROUND STUDY

Machine learning (ML) has emerged as a powerful tool for predicting Alzheimer's disease (AD) by analyzing various types of data, including medical images, genetic information, and cognitive assessments [20]. Researchers have developed several ML approaches to identify patterns and biomarkers associated with AD, enabling early detection and intervention. A systematic review can be found in [4]. Here we outline few recent approaches as follows:

1) Structural MRI-based prediction [19], [21]: Structural magnetic resonance imaging (sMRI) has emerged as a potent tool for computer-aided diagnosis (CAD) of neurological conditions, such as dementia. While convolutional neural networks (CNNs) have demonstrated promise in diagnosing Alzheimer's disease (AD) through learning atrophy patterns from sMRI data, their effectiveness is hampered by the necessity to identify discriminative landmark (LM) positions, potentially impeding overall performance [19]. To overcome this limitation, a novel approach called three-dimensional Jacobian domain convolutional neural network (JD-CNN) is introduced in [19]. This method achieves outstanding classification accuracy without the need for LM detection. By training on features extracted from sMRI transformed into the Jacobian domain, the JD-CNN surpasses existing techniques when evaluated on data from the ADNI database, representing a significant advancement in AD diagnosis. The reported performance matrices are as accuracy 95.42, sensitivity 96.13, specificity 94.17, and AUC as 97.26. Further, the study described in [21] presents an automated method for Alzheimer's disease detection across three stages: control, mild cognitive impairment, and AD. This method utilizes structural MRI data and analyzes co-occurrence matrices and texture statistical measures from MRI images. Through the application of classical machine learning algorithms, it achieves high classification accuracies (up to 93.3%). Moreover, employing a convolutional neural network yields promising results, with accuracies reaching up to 82.2%. The proposed method demonstrates a 4% improvement over existing techniques in discriminating between all groups, showcasing its potential for early AD detection using MRI.
2) PET imaging-based prediction [22], [23]: Positron emission tomography (PET) scans are pivotal in measuring brain activity by detecting radioactive tracers. Deep learning (DL) algorithms can analyze PET images to discern patterns of glucose metabolism or amyloid-beta deposition, key indicators of Alzheimer's disease (AD). In [22], the authors introduce the Inception-ResNet wrapper model to differentiate between healthy controls (HC), mild cognitive impairment (MCI), and AD using multimodal imaging modalities, including magnetic resonance imaging (MRI) and PET scans. They leverage T1-weighted MR and PET images from individuals aged 42 to 95, encompassing HC, MCI, and AD patients. Employing 3D tissue segmentation and fusing atlas-based segmented MR images with PET images, they apply color space transformation and fusion techniques utilizing Fourier and discrete wavelet transforms. With a training-validation-test split of 60%-20%-20%, various convolutional neural networks are utilized to assess the model. Their findings demonstrate superior classification performance, achieving accuracies of 95.5%, 94.1%, and 95.9% for HC vs MCI, MCI vs AD, and AD vs HC, respectively, surpassing existing methods. This underscores the potential of deep learning approaches for automated classification of healthy and dementia stages using combined MRI and PET modalities. In [23], the authors propose a reconstruction-based self-supervised anomaly detection model to address challenges in acquiring labeled medical data for deep neural networks, particularly in diseases like Alzheimer's. This model integrates both MRI and PET scans, featuring a dual-subnetwork encoder with skip connections for enhanced feature encoding and gradient flow, capturing both local and global features. Additionally, it introduces an entropy-based image conversion method. Evaluation results demonstrate superior performance compared to benchmark models in anomaly detection and classification using an encoder. Furthermore, both

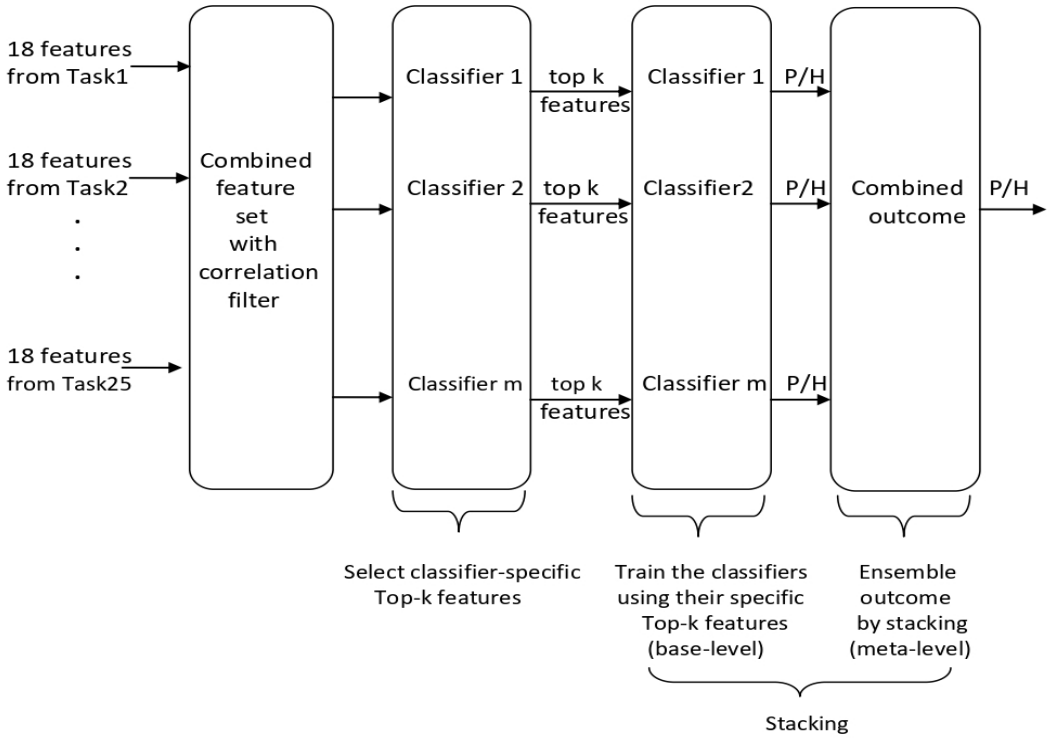


FIGURE 1. The conceptual representation of methodology employed in this study to develop the ensemble predictor for early detection of Alzheimer's Disease (AD) patients. We employed Analysis of Variance(ANOVA) and Recursive Feature Elimination (RFE) to select classifier specific top-k features. "m" is the number of best performing base-level classifiers. "P" and "H" stands for Patients and Healthy people.

supervised and unsupervised models benefit from training with data preprocessed using this approach. The reported performance metrics are as follows: Precision 97.71, Recall 95.9, F1-score 96.37, and Accuracy 97.24.

3) Genetic data analysis [24], [25], [26]: Genetic factors also play a significant role in AD susceptibility. The authors in [24] aim to predict and diagnose Alzheimer's disease (AD) in its early stages using single nucleotide polymorphisms (SNPs) as biomarkers. SNPs are common genetic variations associated with diseases like AD. The study proposes a framework combining machine learning techniques with two feature selection methods: information gain filter and Boruta wrapper. Using gradient boosting tree (GBT) on AD genetic data from ADNI-1 and Whole-Genome Sequencing datasets, the system achieves high accuracy (94.87%). Boruta wrapper feature selection proves superior to the information gain filter, making the proposed system effective for early AD detection. The study in [25] explores Alzheimer's disease (AD) using structural MRI and transcriptome data from the Alzheimer's Disease Neuroimaging Initiative database. It introduces a diagnostic information fusion algorithm, enhancing correlation performance among samples by adding structural constraints to brain regions of interest. Results reveal correlations between genetic variations and brain structure, identifying significant regions affected by multiple risk genes. The study validates the diagnostic significance of these findings for AD. The study in [26] focuses on identifying single nucleotide polymorphisms (SNPs) biomarkers associated with Alzheimer's disease (AD) to improve its classification accuracy. Utilizing deep transfer learning, convolutional neural networks (CNNs) are trained on genome-wide association studies (GWAS) datasets from the AD neuroimaging initiative. Subsequently, deep transfer learning is applied to further train the CNNs on a different AD GWAS dataset to extract final features. These features are then input into a Support Vector Machine for AD classification. Through extensive experiments with multiple datasets and configurations, the study achieves a significant accuracy improvement of 89%, surpassing existing works in the field.

4) Cognitive assessment data analysis [27], [28]: Cognitive tests assess memory, language, and other cognitive functions affected by AD. ML algorithms can analyze cognitive test scores to identify patterns that predict AD progression. The study in [27] explored EEG's efficacy in diagnosing Alzheimer's disease (AD) and mild cognitive impairment (MCI). EEG biomarkers accurately classified participants into healthy, MCI, and AD groups, achieving over 70% accuracy. Notably,

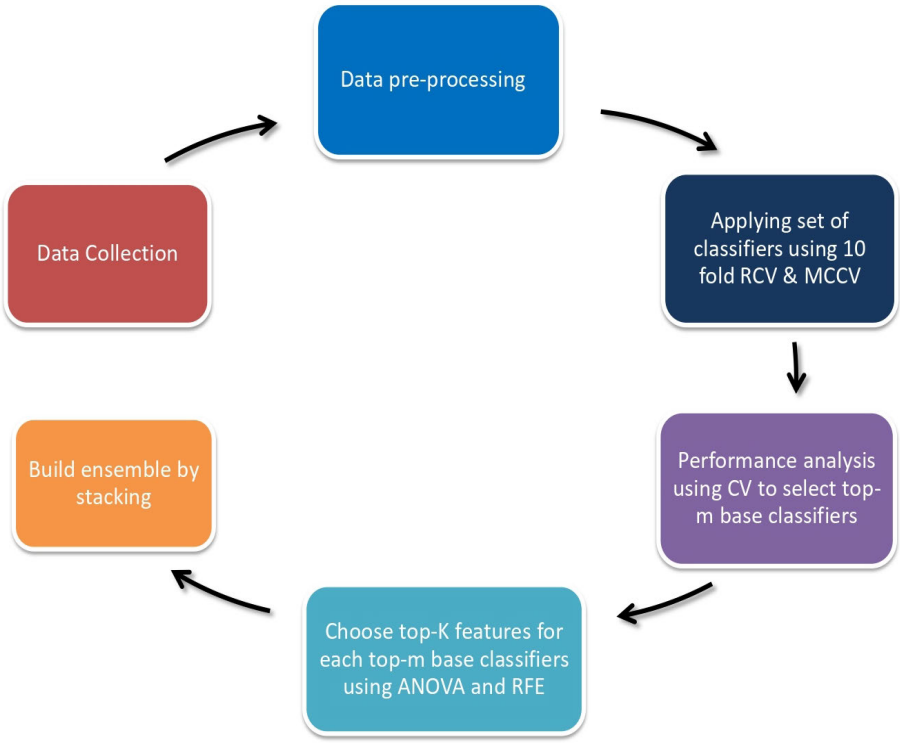


FIGURE 2. Approaches employed are: Data preprocessing encompasses the removal of null values, identification of outliers through Interquartile Range (IQR), scaling via Z-score normalization, and correlation detection using the Pearson correlation coefficient. Performance evaluation includes Repeated Cross Validation (RCV) and Monte Carlo Cross Validation (MCCV) by using ten state-of-the-art performance metrics used in machine learning. Coefficient of Variation (CV) is employed to identify consistent set of classifiers. The selection of the top-k features involves Analysis of Variance (ANOVA) and Recursive Feature Elimination (RFE). The ensemble technique employed is stacking. Note: "m" is the number of best performing (top) base-level classifiers.

EEG features, particularly in parieto-occipital regions, surpassed CSF and APOE biomarkers in predicting disease onset and progression. This underscores EEG's potential for early AD detection and monitoring. Further the study in [28] aimed to develop a predictive model for identifying high-risk individuals of cognitive impairment among older Chinese inpatients. They enrolled 1300 inpatients aged 60 or above and established the model in a developing cohort of 1100 participants, testing it in a validating cohort of 200. Logistic regression analyses identified age, diabetes, depression, and low education as independent factors associated with cognitive impairment. The predictive model, incorporating these variables, yielded a probability of cognitive impairment for each patient. It demonstrated reliability in identifying high-risk individuals (area under curve = 0.790) with high sensitivity (86.2%) but relatively low specificity (59.4%). The model's utility lies in recognizing individuals at high risk rather than ruling out those at low risk of cognitive impairment.

5) Multimodal data fusion [29]: The study in [29] presents a deep learning framework capable of sequentially diagnosing normal cognition, mild cognitive impairment, Alzheimer's disease, and non-AD dementias. The proposed models, which integrate various clinical data, perform comparably to neurologists and neuroradiologists in terms of diagnostic accuracy. By employing interpretability methods, the authors show that their models detect disease-specific patterns that closely align with brain degenerative changes observed during autopsies. Overall, the work highlights the potential of computational approaches in medical diagnosis validation.

Nevertheless, opting for handwriting-based analysis offers several advantages compared to the approaches mentioned. This form of cognitive assessment provides a more direct measure of brain function, capturing the individual's distinctive motor patterns [17]. Moreover, it proves to be a cost-effective and non-invasive alternative to other methods like MRI or PET scans, making it particularly compelling in low- and middle-income countries, which account for 60% of global Alzheimer's Disease incidents. Furthermore, handwriting analysis extends its utility to assess a broader spectrum of cognitive functions, encompassing attention, memory, and executive function.

TABLE 1. Summary of studies on Alzheimer's disease prediction using DARWIN dataset.

Study	ML Model Used	Feature/Data Modality	Techniques Used	Accuracy
De Gregorio et al. [35]	RF, KNN, LDA, GNB, SVM	Task specific	Ensemble multiple task-specific classifiers build upon single type of ML model	91%
Cilia et al. [17]	RF, LR, KNN, LDA, GNB, SVM, DT, MLP, LVQ	Task specific	Ensemble multiple task-specific classifiers build upon top performing and different types of ML model	94.28%
Parziale et al. [38]	Negative Selection Algorithm, Isolation Forest, One-Class SVM	All features	One class classifiers are utilized	97%
Subha et al. [41]	LR, KNN, SVM, DT, RF and AdaBoost ML with Swarm Intelligence	All features	Swarm Intelligence based feature selection are employed with the ML models	90%
Gattulli et al. [37]	RF, LR, KNN, LDA, SVM, BN, GNB, MP, LVQ	Task specific	Mixed tasks of different level of complexity	88%
Onder et al. [39]	XGBoost, GradientBoost, AdaBoost	All features	Categorization methods are employed	85%
Hakan et al. [40]	Ensemble of LGBM, CatBoost, AdaBoost	All features	Hard Ensemble of the employed models	97%
Erdogmus et al. [36]	CNN	1D Features are converted into 2D image	Pre-trained models are utilized for training the classifiers with the constructed 2D images	90%

B. ORIGIN OF THE WORK

Accurate evaluation of handwriting changes indicative of Alzheimer's Disease (AD) necessitates the establishment of rigorous testing criteria. Previous investigations have employed diverse tasks, features, and classifiers for this purpose [4], [30], [31]. However, challenges such as limited datasets and a lack of consensus protocol on feature extraction have posed obstacles. Recognizing the significance of a standardized protocol, Cilia et al. introduced a comprehensive approach in 2018 [32]. Their protocol for collecting handwriting data recommends specific features for extraction, using knowledge from motor control and neuroscience. This approach delves into the intricate dynamics of how distinct brain regions contribute to both handwriting and drawing tasks [32], elucidating how deviations in these regions manifest in overall task performance [33]. Encompassing 25 tasks targeting specific brain regions and covering both handwriting and drawing movements, the protocol meticulously describes the execution of each task using 18 unique features. This collaborative effort culminated in the creation of the DARWIN dataset (Diagnosis AlzheimeR WIth handwriting), representing the largest publicly available comprehensive and invaluable resource for applying machine learning techniques to analyze handwriting data in the context of AD diagnosis.

Expanding on the earlier investigations by Cilia et al. [17], [34], and [42], the current study recognizes the opportunity to enhance the existing Alzheimer's Disease (AD) prediction model using handwriting analysis employing the DARWIN dataset. In De Gregorio et al.'s study [34], the authors trained classifiers utilizing task-specific features, yielding multiple predictors of Alzheimer's Disease (AD) based on specific tasks. Subsequently, they amalgamated the best-performing task-specific predictor to generate the final prediction. The construction of task-specific predictors involved the utilization of a single type of machine learning algorithm. The reported performance state that they are achieved an overall accuracy of 91% with a sensitivity of 83% and a specificity of 100%. In contrast, Cilia et al. [17] employed various machine learning algorithms to create task-specific models, combining the best-performing ones for the ultimate prediction (conceptual diagrams illustrating their methodologies are provided in Supplementary Figures S1 and S2). They achieved highest accuracy of 94.28%, 88.24% specificity and 100% sensitivity. However, both methodologies treated the set of task-specific features as a singular entity, neglecting the potential existence of correlations among these features. In this study we thus aim to address these limitations by prioritizing the selection of crucial features from all tasks to construct a more robust predictor.

Further in [36], using the DARWIN dataset, diverse tasks were assessed using multiple classification models. The key finding emphasizes the need to consider task type and complexity for effective discrimination between healthy individuals and those with Alzheimer's Disease. Their reported accuracy is 83%, specificity of 86% and sensitivity of 81%.

On the other hand in [37] the authors have assessed the performance of three one-class classifier models namely the Negative Selection Algorithm, the Isolation Forest and the One-Class Support Vector Machine, on the DARWIN dataset and achieved impressive accuracy of 97.12, sensitivity of 94.23% and specificity of 100.00% with random Negative Selection Algorithm. However, one major drawback of employing a one-class classifier in medical diagnosis with a small dataset, like DARWIN, is the inherent difficulty in adequately representing the complexity of diverse pathological conditions, as the model relies on limited positive instances AD patients. This limitation contrasts with binary classifiers that benefit from a more balanced representation

of both positive and negative cases, enhancing their ability to discern subtle patterns and generalize to new instances in the context of medical diagnoses. Additionally, the imbalanced data distribution in one-class classification may lead to biased models, impacting the classifier's sensitivity to rare medical conditions, which is crucial in healthcare applications.

In [38], the authors conducted a comprehensive study to diagnose AD using four different categorization methods. These methods included XGBoost, GradientBoost, AdaBoost, and voting classification algorithms. The highest achieved accuracy obtained was 85% by the XGBoost classifiers.

In [39], a tipple ensemble machine learning model is developed and applied for AD detection, by using Light Gradient Boosting Machine, Categorical Boosting, and Adaptive Boosting machine learning classification algorithms were combined with a Hard Voting Classifier. The reported results of the experimental studies by the proposed Ensemble methodology achieved 97.14% Acc, 95% Prec, 100% Recall, 90.25% Spec, and 97.44% F1-score.

In [40], a hybrid ML models with Swarm Intelligence (SI) based feature selection was developed for detection of Alzheimer's disease. They employed several machine learning models, specifically LR, KNN, SVM, DT, RF and AdaBoost, achieving highest accuracy of 90%, precision of 88%, recall of 92%, F1-score 90% and AUC-ROC score 90% with RF and AdaBoost classifier.

In [35], Convolutional Neural Network (CNN) is employed on the DARWIN dataset designed for AD detection through handwriting. They have extracted 2D features from the original 1D dataset and applied to their proposed model. Training and evaluation on this 2D dataset resulted 90.4% accuracy for their proposed model. This suggests the potential of deep learning approach for early and effective AD diagnosis. Table 1 presents the summary of the methods targeting the DARWIN dataset for Alzheimer's disease detection.

In this investigation we delve into the DARWIN dataset, resulting in the creation of a highly precise ensemble classifier based on stacking technique designed for the early detection of Alzheimer's Disease (AD) cases. The predictive model is a fusion of diverse machine learning models, each meticulously trained with its own distinct set of top-ranking features across all 25 tasks (450 features). To pinpoint the most effective base-level classifiers, we leverage both Repeated 10-fold cross-validation (RCCV) and Monte Carlo cross-validation (MCCV) techniques. We compute the mean values along with standard deviations for the ten performance metrics: accuracy, sensitivity, specificity, precision, True Positive Rate, False Positive Rate, F1-score, Matthews Correlation Coefficient, Cohen Cappa and Area Under the Receiver Operating Characteristic Curve by using RCCV and MCCV approaches. The coefficient of variation, obtained by dividing the standard deviation by the mean for each metric, serves as the basis for ranking the classifiers specific to that metric. These metric-specific rankings are then consolidated to establish a global ranking, offering a comprehensive evaluation of the classifiers across all ten performance metrics.

By employing Analysis of Variance (ANOVA) and Recursive Feature Elimination (when applicable), we identify the top-k features unique to each classifier for the best-performing base classifiers. Subsequently, the chosen base classifiers undergo additional training using their specific top-k features and are amalgamated through a stacking ensemble technique. This final ensemble classification approach yields predictive outcomes for identifying individuals at risk of Alzheimer's disease. The resultant feature-specific multi-classifier ensemble model showcases remarkable performance metrics, boasting a 97.14% accuracy, 95% sensitivity, 100% specificity, 100% precision, 95% True Positive Rate (TPR), 0% False Positive Rate (FPR), 97.4% F1-score, 94.37% Matthews Correlation Coefficient (MCC), Cohen Kappa of 94.21%, and an impressive 97.5% Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Furthermore, comparative performance analysis with the state-of-the-art models targeting DARWIN dataset shows superiority of the proposed ensemble model.

The methodology employed in this study is detailed in Figure 1, providing a visual representation of the steps undertaken in our analysis. The rest of the paper is organized as follows: Experimental setup is described in Section II; While Section III contains the Results and Discussion; Conclusion and Future work are given in Section IV.

TABLE 2. Average demographic data of participants. Standard deviations are shown in parentheses.

Participant class	Age	#Women	#Men
Patients	71.5 (9.5)	46	44
Healthy people	68.9 (12)	51	39

II. EXPERIMENTAL SETUP

This study employed a multi-faceted experimental methodology encompassing several key stages. Firstly, the process involved comprehensive data collection to assemble a relevant dataset. Subsequently, a meticulous data pre-processing phase was implemented to refine and prepare the collected data. The next step entailed the selection of base-level classifiers, achieved through performance analysis utilizing both Repeated Cross-Validation (RCV) and Monte Carlo Cross-Validation (MCCV) techniques. Following this, the study identified the top-k features from the best-performing base-level classifiers to focus on the most influential features. Lastly, an ensemble model was constructed by stacking the chosen base-level classifiers. To enhance clarity and understanding, a visual representation of these approaches is presented in Figure 2. This figure serves the purpose of providing a clear and concise visualization of the sequential steps undertaken in the experimental process.

A. DATA COLLECTION

The DARWIN (Diagnosis AlzheimeR WIth haNdwriting) dataset has been meticulously curated to support the early detection of Alzheimer's disease by examining handwriting features. This valuable dataset includes data from 174 individuals, comprising 89 persons diagnosed with Alzheimer's disease and 85 healthy participants. This dataset was generated by combining 25 unique tasks (described in Supplementary Table S1) of three categories: Memory and dictation (M), Graphic (G), and Copy (C). Each of the task is providing 18 distinct handwriting features (described in Supplementary Table S2). This compilation resulted in an extensive collection of 450 features that pertain to handwriting analysis. Average demographic information about the participants for creating the dataset is presented in Table 2.

B. DATA PREPROCESSING

Data preprocessing is vital in data analysis and machine learning, involving the refining and structuring of raw data to ensure proper formatting for subsequent stages like model training. We initially apply null value filtering from the dataset. Let, x be a data point in a dataset D, and $isNull(x)$ be an indicator function that equals 1 for null values and 0 for non-null values, then the filtering of null values can be expressed as:

$$D' = \{x \in D | isNull(x) = 0\} \tag{1}$$

This notation specifies that D' contains all data points x in the dataset D where $isNull(x)$ equals 0 (indicating x is not null).

The outliers are the data points that significantly deviate from the dataset's typical distribution, either with exceptionally high or low values. The Interquartile Range (IQR) is a statistical measure widely employed in outlier detection. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1) of a dataset. Mathematically, the IQR is defined as:

$$IQR = Q3 - Q1 \tag{2}$$

To identify potential outliers, a common criterion is to consider data points that fall below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR as outliers. The IQR is a robust and non-parametric method, making it a valuable tool in outlier detection tasks, especially when the data distribution is not well-known.

Scaling or normalization ensures consistent and comparable feature scales, which is vital for machine learning algorithms' convergence, performance, and interpretability. We employed standardization (Z-score normalization) which can be defined as for a given data point ''x'' in a dataset D is as follows:

$$Z = (x - \mu)/\sigma \tag{3}$$

where, Z is the standardized value (Z-score) of the data point, x is the original data point, μ is the mean (average) of the dataset and σ is the standard deviation of the dataset. This process results in transforming the data into a distribution with a mean of 0 and a standard deviation of 1.

Finally, identifying highly correlated or redundant features is crucial, as such features can negatively impact model performance. To perform correlation-based feature filtering on a dataset D, we employ the Pearson correlation coefficient (corr), which measures the linear relationship between variables. Let X and Y represent two variables within D, and the formula for calculating the Pearson correlation coefficient is given as:

$$\text{corr}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{4}$$

where, $\text{corr}(X, Y)$ is the Pearson correlation coefficient between variables X and Y. X_i and Y_i are individual data points in the X and Y, respectively. $\bar{X}$ and $\bar{Y}$ are the mean values of the X and Y, respectively, n is the number of data points in the datasets. To filter features based on correlation within D, we can specify a correlation threshold, denoted as correlation_threshold. Features X_k in D are selected based on the condition:

$$D_{\text{filtered}} = \{X_k \in D \,|\, \forall X_i, X_j \in D, i \neq j : |\text{corr}(X_i, X_k)| \leq \text{correlation\_threshold}\} \tag{5}$$

This notation represents the dataset D_{filtered} containing features X_k for which the absolute value of the Pearson correlation coefficient between X_k and any other feature X_i (where $i \neq k$) is less than or equal to the specified correlation threshold. This procedure allows for effective feature selection based on correlation within the dataset. The complete road-map for data pre-processing is presented in Table 3.

C. PERFORMANCE EVALUATION METRICS

In order to create an efficient predictor for Alzheimer's disease (AD), we conducted a comprehensive assessment of ten state-of-the-art machine learning classification algorithms by using two distinct cross-validation approaches: Repeated 10-fold cross-validation (RCV) and Monte Carlo Cross-validation (MCCV). These algorithms encompass Random Forest (RF), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Gaussian Naive Bayes (GNB), Extra Trees (ET), XGBoost (XGB), k-Nearest Neighbors (KNN), Support Vector Machines (SVM), Multi-Layer Perceptron (MLP) and Decision Tree (DT). To identify the most precise set of classifiers we analyze and compare the scores of ten widely used performance metrics in machine learning: accuracy, sensitivity, specificity, precision, True Positive Rate, False Positive Rate, Cohen's Kappa, F1 score, Matthews Correlation Coefficient (MCC) and AUC-ROC. Additionally, we employed the coefficient of variation to assess the consistency of these scores. Consistency is of paramount importance in medical applications, ensuring not only patient safety but also fostering confidence in healthcare

TABLE 3. Data Pre-processing Steps. Note: The following pre-processing steps are applied in-order on the DARWIN dataset.

Pre-processing Step in-order	Purpose	Techniques Used	Notes
1. Null Value Removal	Remove missing values from the dataset	Deletion, Imputation	This dataset don't have missing values.
2. Outlier Removal	Identify and handle outliers in the dataset	IQR method	Outliers replaced with mode value
3. Z-score Normalization	Standardize the scale of numerical features	Z-score Normalization	Ensure consistent range for numerical features
4. Removal of Highly Correlated Features	Eliminate features with high correlation	Pearson's Correlation Analysis	Remove redundant or collinear features

TABLE 4. Structure and components of a typical confusion matrix.

	Predicted	
Actual	Positive	Negative
Positive	TP	FN
Negative	FP	TN

AI systems, maintaining ethical and regulatory compliance, and optimizing the efficient allocation of healthcare resources. In the subsequent section, we present a concise overview of these performance evaluation metrics and the cross-validation approaches employed in our study.

All the performance metrics in machine learning is based on the concept of confusion matrix which is a table (Table 4) used to assess the performance of a classification algorithm. The main components of a confusion matrix are

1) True Positive (TP): The number of instances correctly predicted as positive.
2) True Negative (TN): The number of instances correctly predicted as negative.
3) False Positive (FP): The number of instances incorrectly predicted as positive.
4) False Negative (FN): The number of instances incorrectly predicted as negative.

1) ACCURACY

Accuracy is a measure of overall correctness in a diagnostic test or predictor model. It can be defined as

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{6}$$

2) SENSITIVITY AND SPECIFICITY

Sensitivity measures the proportion of true positives (patients) correctly classified, while specificity measures the proportion of true negatives (healthy) correctly classified.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{7}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

3) PRECISION

Precision measures the proportion of true positives among all predicted positives, emphasizing the classifier's ability to avoid false positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

4) TRUE POSITIVE RATE (TPR) AND FALSE POSITIVE RATE (FPR)

The True Positive Rate (TPR), also known as *Sensitivity or Recall*, signifies the proportion of actual positive cases correctly identified by the classifier. On the other hand, the False Positive Rate (FPR) measures the proportion of actual negative cases incorrectly identified as positive.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{11}$$

5) COHEN'S KAPPA

Cohen's Kappa measures of agreement between the model's predictions and the gold standard diagnoses

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{12}$$

where: - P_o is the observed agreement between the classifier's predictions and the actual diagnoses. - P_e is the expected agreement by chance, calculated based on the marginal probabilities of agreement for the classifier and the actual diagnoses.

6) F1 SCORE

The F1 score holds significant importance in medical diagnosis due to its ability to strike a crucial balance between precision and recall, can be defined as

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

7) MATTHEWS CORRELATION COEFFICIENT (MCC)

The Matthews Correlation Coefficient (MCC) considers true positives, true negatives, false positives, and false negatives, thereby capturing the full spectrum of diagnostic accuracy. It ranges from -1 (perfect disagreement) to 1 (perfect agreement), and can be defined as

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{14}$$

8) ROC-AUC

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) holds crucial importance in medical diagnosis by offering a comprehensive assessment of classification model performance. Its threshold-independence ensures that it evaluates a model's ability to discriminate between classes across all possible decision thresholds, making it invaluable in healthcare settings where optimal threshold selection can vary. A high AUC-ROC signifies strong discriminative power, allowing for robust performance assessment, model comparison, and the selection of diagnostic tools that reliably differentiate between individuals with and without

a condition. In turn, this contributes to early detection, informed clinical decisions, and improved patient outcomes in medical practice.

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}) \tag{15}$$

where, TPR (True Positive Rate) is also known as sensitivity or recall, and it is plotted on the y-axis, and FPR (False Positive Rate) is plotted on the x-axis.

9) CPU TIME AND MEMORY REQUIREMENT

We consider both the CPU time and memory requirement of the implementation for the proposed model. CPU time refers to the duration the CPU spends executing instructions for a specific task, crucial for assessing computational efficiency. In Python, this can be measured using the 'time' or 'timeit' modules. Memory requirements denote the RAM consumed during program execution, vital for optimizing resource usage. Python libraries like 'memory_profiler' facilitate memory measurement, aiding in efficient memory management for machine learning tasks.

D. CROSS VALIDATION TECHNIQUES

1) REPEATED K-FOLD CROSS VALIDATION

Repeated Cross-Validation is a technique used to assess the consistency in performance of a machine learning model by repeatedly splitting the dataset into training and testing subsets, ensuring robustness or stability. It is defined as follows:

Data Splitting:

- Divide the dataset into k subsets (folds) of equal size, denoted as $D_1, D_2, \ldots, D_k$.

Repeated Cross-Validation Process:

- Repeat the following process R times:
  1) Randomly shuffle the dataset to create variability in data splits: $ShuffledData = Shuffle(Data)$.
  2) For each repetition r ($1 \leq r \leq R$), do the following:
     a) **k-Fold Split**:
        - Divide the shuffled data into k folds: $D_1^r, D_2^r, \ldots, D_k^r$.
     b) **Model Training and Testing**:
        - Train the model on $k-1$ folds: $Model_i^r = Train(ShuffledData - D_i^r)$.
        - Evaluate the model on fold i to compute a performance metric: $Metric_i^r = Evaluate(Model_i^r, D_i^r)$.
     c) **Performance Aggregation**:
        - Calculate the average performance metric for this repetition: $Avg\_Metric^r = \frac{1}{k}\sum_{i=1}^{k} Metric_i^r$.

Performance Estimation:

- After R repetitions, we'll have R sets of average performance metrics: $\{Avg\_Metric^1, Avg\_Metric^2, \ldots, Avg\_Metric^R\}$.
- Estimate the overall model performance by calculating:
  - Mean performance: $\mu = \frac{1}{R}\sum_{r=1}^{R} Avg\_Metric^r$.
  - Standard deviation of performance: $\sigma = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(Avg\_Metric^r - \mu)^2}$.

In Supplementary Figure S3, we can observe the process of 10-fold cross-validation. In the case of a repeated or nested approach, this entire procedure is reiterated a predefined number of times to derive a comprehensive average score for the performance metrics. In our study, we employed repeated 10-fold cross validation i.e.; setting k = 10.

2) MONTE CARLO CROSS-VALIDATION

Monte Carlo Cross-Validation (MCCV) is a technique, depicted in Supplementary Figure S4, for assessing the performance of a machine learning model by repeatedly randomizing the data and splitting it into training and testing sets. We repeat the process 100 times, i.e.; we set R = 100 (parameter defined below). This method is particularly useful when traditional cross-validation approaches are impractical due to limited data or when the data is not naturally divided into folds.

MCCV Process:

- Define the number of repetitions as R.
- For each repetition r ($1 \leq r \leq R$), perform the following steps:
  1) Random Shuffling: Randomly shuffle the entire dataset to create variability in data splits: $ShuffledData = Shuffle(Data)$.
  2) Data Splitting: Split the shuffled data into training and testing sets. The size and partitioning can vary with each repetition.
  3) Model Training: Train the machine learning model on the training set: $Model^r = Train(TrainingData^r)$.
  4) Model Testing: Evaluate the model on the testing set to compute a performance metric: $Metric^r = Evaluate(Model^r, TestingData^r)$.

Performance Estimation:

- After R repetitions, you'll have R performance metrics: $\{Metric^1, Metric^2, \ldots, Metric^R\}$.
- Estimate the overall model performance by calculating summary statistics:
  - Mean performance: $\mu = \frac{1}{R}\sum_{r=1}^{R} Metric^r$.
  - Standard deviation of performance: $\sigma = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(Metric^r - \mu)^2}$.

The repeated 10-fold cross-validation yields unbiased results but exhibits high variance, while Monte Carlo cross-validation produces high bias but low variance, leading to nearly opposite outcomes. Hence, we have embraced a strategy that incorporates two divergent approaches, utilizing classifiers commonly recommended by both methods as the foundational classifiers. This approach aims to enhance the robustness of our selection process.

E. SUPERVISED LEARNING ALGORITHMS

In supervised machine learning, the initial step involves using a labeled training dataset to create models that can recognize patterns within the data. These well-trained models are subsequently applied to an unlabeled testing dataset to classify its contents into relevant categories. The following subsection provides a brief overview of various supervised machine learning algorithms used for early detection of Alzheimer's disease.

Decision Tree (DT): A Decision Tree recursively splits the data based on the most significant attribute. No specific mathematical equation for Decision Tree itself, but it uses criteria like Gini impurity or entropy for splitting.

Random Forest (RF): Random Forest is an ensemble of decision trees. The final prediction is obtained by averaging or taking a majority vote from the individual trees.

Extra Trees (Extra Tree): Extra Trees, like Random Forest, is an ensemble of decision trees but with more randomness in feature selection and splitting. It uses the same mathematical rules for splitting nodes as traditional decision trees.

Logistic Regression (LR): Logistic Regression models the probability of an instance belonging to a particular class using the logistic function (sigmoid function).

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}} \tag{16}$$

Linear Discriminant Analysis (LDA): LDA finds the linear combinations of features that maximize the separation between different classes. Mathematical Equation for Linear Discriminant:

$$\mathbf{y} = \mathbf{W}^T\mathbf{x} \tag{17}$$

where $\mathbf{W}$ is the weight vector and $\mathbf{y}$ is the linear discriminant.

Gaussian Naive Bayes (GNB): GNB is based on Bayes' theorem and assumes that features are conditionally independent given the class. Mathematical Equation for Naive Bayes:

$$P(y|\mathbf{x}) \propto P(y)\prod_{i=1}^{n} P(x_i|y) \tag{18}$$

XGBoost (XGB): XGBoost is a gradient boosting algorithm that minimizes a loss function by adding decision trees iteratively. Objective Function in XGBoost:

$$\text{Obj}(\Theta) = \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{19}$$

k-Nearest Neighbors (KNN): KNN assigns a class label to an instance based on the majority class among its k nearest neighbors. No specific mathematical equation for KNN itself, but it uses distance metrics like Euclidean or Manhattan distance.

Support Vector Machine (SVM): SVM finds the hyperplane that best separates different classes with the maximum margin. Mathematical Equation for Linear SVM:

$$\text{maximize } \frac{2}{\|\mathbf{w}\|} \text{ subject to } y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 \quad \text{for } i = 1, \ldots, n \tag{20}$$

Multi-Layer Perceptron (MLP): MLP is a type of artificial neural network with multiple layers of interconnected nodes, including input, hidden, and output layers. Mathematical Equations for Forward Propagation:

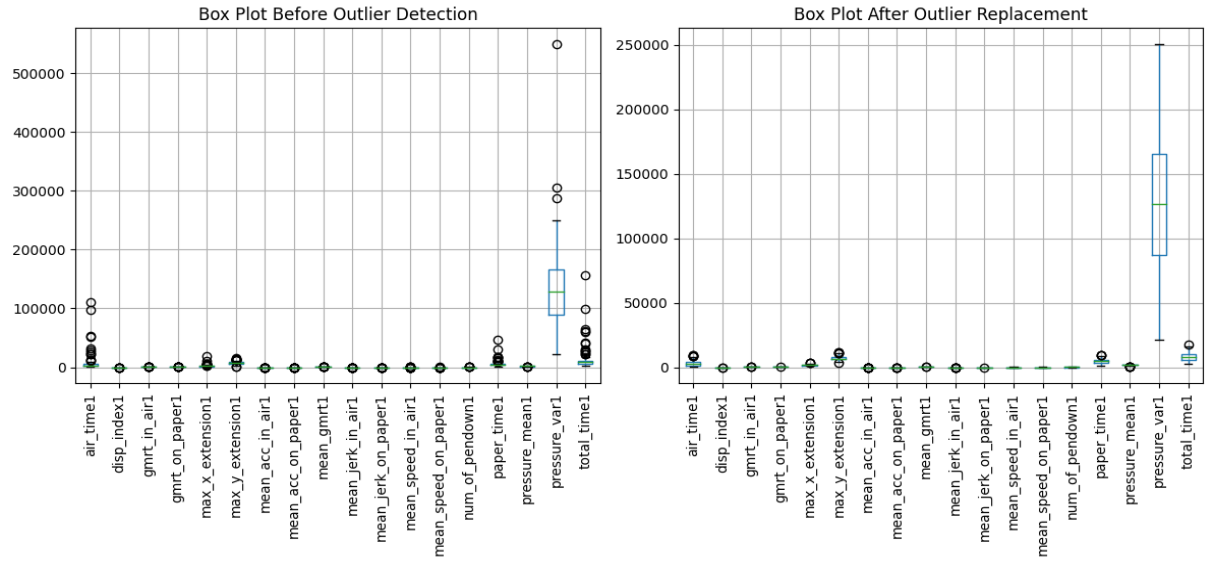
$$z_j^{(l)} = \sum_i w_{ji}^{(l)} a_i^{(l-1)} \tag{21}$$

$$a_j^{(l)} = \sigma(z_j^{(l)}) \tag{22}$$

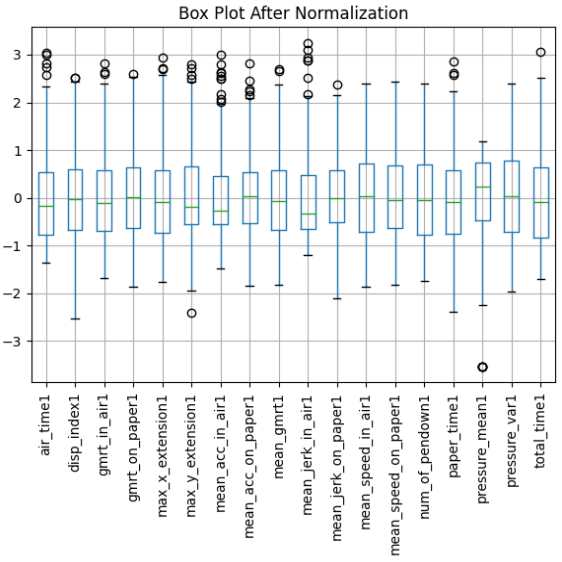
F. SELECTING TOP-K FEATURES

In the realm of feature selection, the "selectBestK" function from scikit-learn (often denoted as "sklearn") stands out as a widely utilized machine learning library in Python. This function is instrumental in pinpointing and preserving the most informative features within a given dataset. Specifically, it leverages Analysis of Variance (ANOVA) as a criterion to gauge the relevance of each feature with respect to the target variable. ANOVA evaluates variance among different groups (classes) in a dataset, effectively discerning features that significantly contribute to the variation in the target variable. By systematically evaluating all features, the selectBestK function identifies the top k features with the highest ANOVA scores, signifying their robust association with the target variable. This approach proves invaluable in machine learning and statistical modeling, enabling practitioners to concentrate on a subset of features likely to meaningfully enhance a model's predictive power, thereby improving efficiency and interpretability. The incorporation of ANOVA within the selectBestK function empowers practitioners to make informed decisions about feature inclusion, optimizing machine learning workflows.

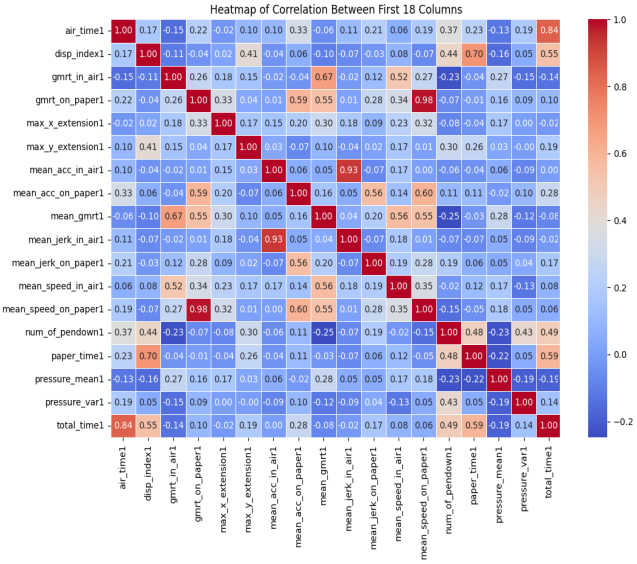
Conversely, Recursive Feature Elimination (RFE) represents another noteworthy feature selection technique in machine learning. RFE systematically assesses and ranks feature importance through iterative model training, progressively discarding the least significant features. Commencing with the entire feature set, RFE assigns importance scores to each feature, eliminates the least important ones, and continues this process until a specified number of features or optimal performance is attained. This iterative strategy aids in identifying a subset of features that wield the most influence over a model's predictive power, leading to enhanced model efficiency, interpretability, and often superior performance. In our feature selection endeavors for our set of classifiers, we judiciously applied RFE when appropriate; otherwise, we relied on the selectBestK method to achieve an optimal balance between thorough feature assessment and computational efficiency.



(a) Outlier detection and replacement of the Task 1 feature set.



(b) Box plot after normalization of the Task 1 feature set.



(c) Heat map of the Task 1 feature set.

FIGURE 3. Exploratory Data Analysis (EDA) of Feature Set for Task-1, including (a) Outlier Detection and Replacement (b) Scaling and (c) Heat map. X-axis in figures (a), (b) and (c) represents the feature names, while Y-axis represents the values for that feature. Note that after normalization the range of feature values are within −1 to +1.

G. PROPOSED MODEL

We developed a novel prognostic model utilizing the stacking ensemble technique to identify cases with the onset of Alzheimer's disease. Stacking is an ensemble learning approach that amalgamates multiple base models to construct a meta-model, enabling predictions based on the collective outputs of the base models. This approach capitalizes on the individual strengths of diverse base models, allowing them to complement each other and mitigate weaknesses. Our model creation process involves three primary steps: (i) training the base models with their distinctive top-k features, (ii) training the meta-model, and (iii) predicting novel instances.

1) TRAINING BASE CLASSIFIERS WITH MODEL SPECIFIC TOP-K FEATURES:

Let we have M base models denoted as $f_1, f_2, \ldots, f_M$. Each base model f_i undergoes independent training on the dedicated training dataset $(X^{f_i}_{\text{train}}, y_{\text{train}})$ utilizing a model-specific set of top-k features meticulously chosen for the model f_i. For a given input sample x_i, each base model produces a prediction $h_j(x_i)$, where j represents the index of the base model. The predictions from all base models form a vector:

$$H_i = [h_1(x_i), h_2(x_i), \ldots, h_M(x_i)] \tag{23}$$

Each $h_j(x_i)$ is generated using the pre-selected top-k features for the j-th base model.

2) TRAINING META-MODEL WITH WEIGHTED PREDICTIONS:

A meta-model, often called the stacking model, is trained using the predictions of the base models as features. The training dataset for the meta-model consists of the predictions $\{H_i\}$ obtained from the base models and the corresponding true labels y_{train}. Let $F(x)$ represent the meta-model, which takes the predictions H_i as input and produces the final ensemble prediction. The meta-model is trained using a training dataset $(H_{\text{train}}, y_{\text{train}})$, where H_{train} is a matrix with each row corresponding to the predictions for a training sample.

$$H_{\text{train}} = \begin{bmatrix} H_1 \\ H_2 \\ \vdots \\ H_N \end{bmatrix} \tag{24}$$

The training process involves finding the parameters of the meta-model $F(x)$ that minimize a specified loss or objective function, considering weighted predictions:

$$\min_{\theta,\boldsymbol{\alpha}} \sum_{i=1}^{N} \boldsymbol{\alpha} \cdot \text{Loss}(F(x_i;\theta), y_i) \tag{25}$$

Here, θ represents the parameters of the meta-model, and $\boldsymbol{\alpha}$ is a vector of learned weights assigned to the predictions of each base model based on coefficient of variation score defined in Eq.28.

3) MAKING PREDICTIONS WITH TOP-K FEATURES:

To make predictions on a new, unseen sample x_{test}, the process involves:

a. Each base model produces predictions $H_{\text{test}} = [h_1(x_{\text{test}}), h_2(x_{\text{test}}), \ldots, h_M(x_{\text{test}})]$.

b. The meta-model $F(x)$ takes H_{test} as input and produces the final ensemble prediction $F(x_{\text{test}})$.

$$F(x_{\text{test}}) = F(H_{\text{test}}) \tag{26}$$

Each $h_j(x_{\text{test}})$ is generated using the pre-selected top-k features for the j-th base model.

4) AN ALTERNATIVE MODEL BASED ON SOFT VOTING (MAJORITY VOTING)

The soft voting technique involves combining the predicted probabilities from individual models. For N models, each denoted as M_i, the ensemble prediction $P_{\text{ensemble\_Majority}}$ for a given input $\mathbf{x}$ is obtained by averaging the predicted probabilities:

$$P_{\text{ensemble\_Majority}}(y = c|\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} P_{M_i}(y = c|\mathbf{x}) \tag{27}$$

Here, $P_{M_i}(y = c|\mathbf{x})$ represents the probability predicted by model M_i for the input $\mathbf{x}$ belonging to class c. The final prediction is the class with the highest average probability.

III. RESULTS AND DISCUSSION

Our research endeavors encompassed an extensive array of experiments aimed at developing a robust and accurate predictor for the early detection of Alzheimer's disease. Initially, we applied cutting-edge data pre-processing techniques to optimize the DARWIN dataset, enhancing the accuracy and effectiveness of our predictor. Leveraging a stacking ensemble technique, we amalgamated distinct machine learning classification models, forming a two-tiered structure of classifiers- the base level and the meta-level built upon them (Figure 1).

In the pursuit of identifying the most proficient base-level classifiers, we applied two distinct cross-validation strategies, namely RCV and MCCV. Using these strategies, we generated scores for ten state-of-the-art machine learning performance metrics. The primary ranking of the base-level classifiers relied on the coefficient of variation (CV) of the scores for each performance metric. Subsequently, we integrated these scores to establish the global ranking, incorporating insights from both RCV and MCCV strategies to select the optimal base-level classifiers. After selecting the top-performing base-level classifiers, we utilized ANOVA and RFE methods to select classifier-specific top-k features. Subsequently, we retrained the classifiers using this refined set of features. The retrained base-level classifiers underwent a stacking ensemble technique, where their predictions were merged to create a meta-level classifier. The predictions generated by this meta-level classifier constituted the conclusive stage of the diagnostic results.

All experiments were conducted on a modest computing setup, featuring an Intel(R) Core(TM) i5-5200U CPU @ 2.20 GHz, 4GB RAM, a 1TB HDD and also using computational resource from the freely available version of Google Colab.

A. RESULTS FROM EXPLORATORY DATA ANALYSIS (EDA)

In the exploration of the DARWIN dataset, our analysis delved into its underlying characteristics through exploratory data analysis (EDA). This comprehensive investigation resulted in a series of insights presented in subsequent sections. To begin, we utilized the "isna()" function from the "pandas" library, revealing that the DARWIN dataset contained no missing values. Subsequently, we move on to perform an analysis focused on identifying outliers. Figure 3(a) illustrates the distribution of quantitative features for Task 1 in the dataset, highlighting outliers as points outside the boxes and whiskers. Outliers were identified using the Interquartile Range (IQR) method and replaced with the corresponding "mode" values for each feature. The figure also presents the dataset at post-outlier replacement. Our experimentation extends to the features corresponding to all other 24 tasks in the DARWIN dataset, employing the IQR method to detect and resolve outliers across tasks. After filtering and replacing these outliers, the remaining instances were retained for further examination.

TABLE 5. Scores of performance metrics (mean value ± standard deviation) of the machine learning classification algorithms using repeated 10-fold cross validation.

Classification Algorithm	Acc	Sn (recall)	Sp	Precision	TPR	FPR	Cohen's Kappa	F1	MCC	AUC-ROC
RF	0.86 ± 0.08	0.90 ± 0.10	0.89 ± 0.00	0.84 ± 0.10	0.67 ± 0.47	0.37 ± 0.45	0.71 ± 0.15	0.86 ± 0.07	0.72 ± 0.15	0.85 ± 0.08
LR	0.83 ± 0.08	0.81 ± 0.13	0.89 ± 0.00	0.86 ± 0.10	0.67 ± 0.47	0.37 ± 0.45	0.66 ± 0.17	0.83 ± 0.09	0.67 ± 0.17	0.83 ± 0.08
LDA	0.70 ± 0.12	0.66 ± 0.17	0.89 ± 0.00	0.73 ± 0.13	0.62 ± 0.44	0.37 ± 0.45	0.39 ± 0.23	0.68 ± 0.14	0.41 ± 0.23	0.70 ± 0.12
GNB	0.84 ± 0.08	0.88 ± 0.10	0.78 ± 0.00	0.82 ± 0.10	0.67 ± 0.47	0.41 ± 0.43	0.67 ± 0.17	0.85 ± 0.08	0.68 ± 0.17	0.83 ± 0.09
ExtraTree	0.86 ± 0.07	0.89 ± 0.09	0.78 ± 0.00	0.86 ± 0.10	0.62 ± 0.44	0.41 ± 0.43	0.72 ± 0.15	0.87 ± 0.07	0.73 ± 0.14	0.86 ± 0.07
XGB	0.85 ± 0.08	0.87 ± 0.11	0.89 ± 0.00	0.86 ± 0.10	0.62 ± 0.44	0.37 ± 0.45	0.71 ± 0.15	0.86 ± 0.08	0.72 ± 0.15	0.85 ± 0.08
KNN	0.62 ± 0.08	0.26 ± 0.15	1.00 ± 0.00	0.92 ± 0.28	0.46 ± 0.41	0.33 ± 0.47	0.25 ± 0.15	0.39 ± 0.19	0.36 ± 0.15	0.63 ± 0.07
SVM	0.82 ± 0.09	0.79 ± 0.13	0.89 ± 0.00	0.87 ± 0.10	0.62 ± 0.44	0.37 ± 0.45	0.65 ± 0.17	0.82 ± 0.10	0.66 ± 0.17	0.83 ± 0.09
MLP	0.84 ± 0.08	0.84 ± 0.12	0.89 ± 0.00	0.85 ± 0.10	0.67 ± 0.47	0.37 ± 0.45	0.67 ± 0.16	0.84 ± 0.09	0.68 ± 0.16	0.84 ± 0.08
DT	0.74 ± 0.09	0.73 ± 0.14	0.67 ± 0.00	0.76 ± 0.12	0.62 ± 0.44	0.44 ± 0.42	0.48 ± 0.19	0.74 ± 0.10	0.49 ± 0.19	0.74 ± 0.09

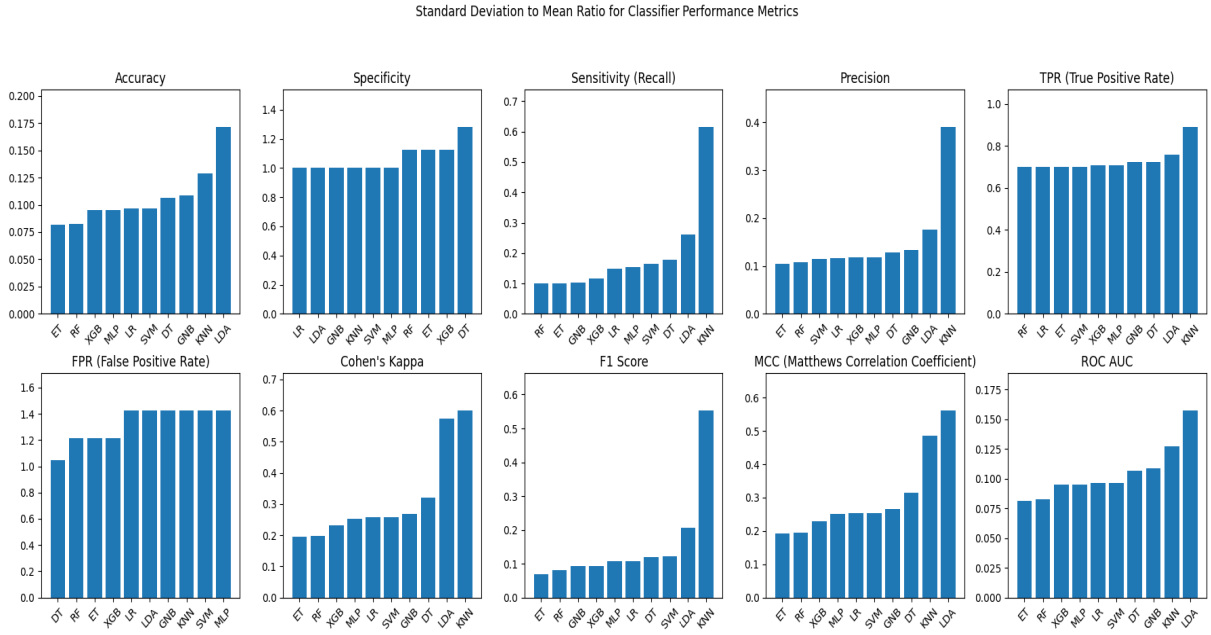


FIGURE 4. Ranking of the base-level classifiers using CV_m scores using RCV approach. Note CV value for specificity is computed by 1/mean value.

In Figure 3(b), the impact of feature scaling on the dataset for Task 1 is depicted through z-score normalization. Similar computations were applied to the features of the remaining 24 tasks. This process emphasized the importance of scaling in aligning feature ranges, preventing features with larger magnitudes from dominating those with smaller magnitudes.

Our analysis further extended to include a comprehensive correlation analysis of all features. The results were presented in a heatmap (for Task 1), displayed in Figure 3(c). This heatmap provided a visual representation of the correlation values between features of Task 1. Negative correlation values denoted an inverse relationship, while a value of zero signified no correlation. Some features demonstrate particularly robust correlations (>90%), and a similar high level of correlation is observed in the feature sets of other tasks. This underscores the necessity for feature selection within the set of features associated with a task. It suggests that treating a task, with its set of features, as a singular entity (as done in Cilia et al. [17]) is likely to include highly correlated features, potentially leading to performance degradation in classification tasks. The omission of highly correlated features across all the tasks resulted in formation of a dataset containing 337 features.

The updated dataset presents a more organized and transparent sequence of operations, offering a clear rationale for each step in the analysis of the DARWIN dataset. Notably, this revised version provides a superior representation of the original dataset by eliminating highly correlated features.

B. RESULTS FROM MACHINE LEARNING ANALYSIS

1) SELECTING BASE-LEVEL CLASSIFIERS WITH REPEATED 10-FOLD CROSS VALIDATION

Each machine learning classification algorithm, as outlined in section 3.4, is individually applied to the pre-processed DARWIN dataset, which comprises 337 features. The analysis employs repeated 10-fold cross-validation (RCV). The primary goal of this experiment is to identify a set of reliable and accurate base-level classifiers by evaluating and comparing the performance metric scores derived from RCV.

The results, encompassing mean (μ) values and standard deviations (ρ) for each performance metric, are detailed in

TABLE 6. Ranking of the classifiers based on CV score of the individual performance metrics using RCV.

Performance Metric	Ranks									
	1	2	3	4	5	6	7	8	9	10
ACC	ET	RF	XGB	MLP	LR	SVM	DT	GNB	KNN	LDA
SP	LR	LDA	GNB	KNN	SVM	MLP	RF	ET	XGB	DT
SN	RF	ET	GNB	XGB	LR	MLP	SVM	DT	LDA	KNN
PRE	ET	RF	SVM	LR	XGB	MLP	DT	GNB	LDA	KNN
TPR	RF	LR	ET	SVM	XGB	MLP	GNB	DT	LDA	KNN
FPR	DT	RF	ET	XGB	LR	LDA	GNB	KNN	SVM	MLP
Cohen's K	ET	RF	XGB	MLP	LR	SVM	GNB	DT	LDA	KNN
F1-score	ET	RF	GNB	XGB	MLP	LR	DT	SVM	LDA	KNN
MCC	ET	RF	XGB	MLP	LR	SVM	GNB	DT	KNN	LDA
AUC-ROC	ET	RF	XGB	MLP	LR	SVM	DT	GNB	KNN	LDA

TABLE 7. Ranking of the classifier based on Global Ranking generated from RCV.

Rank	Classifier	Global Ranking
1	ET	22
2	RF	23
3	LR	43
4	XGB	43
5	MLP	55
6	SVM	60
7	GNB	61
8	DT	71
9	LDA	83
10	KNN	89

Table 5. To assess classifier consistency, the coefficient of variation (CV) is employed. Classifiers with lower CV scores are generally considered more consistent and stable. This characteristic is attributed to their reduced susceptibility to overfitting during training, enhancing their ability to perform effectively on previously unseen data.

Let M be the set of performance metrics, where $|M| = 10$ (in our case). For each metric $m \in M$, there are two values: the mean μ_m and the standard deviation σ_m. The Coefficient of Variation (CV) for each metric m is calculated as the ratio of the standard deviation to the mean as

$$CV_m = \frac{\sigma_m}{\mu_m} \tag{28}$$

Figure 4(a) shows the CV_m for all the classifiers in ascending order.

For each classifier, we calculate the sum of the CV values across all metrics. Let C be the set of classifiers, where $|C| = 10$ (in our case). For each classifier $c \in C$, calculate the sum of CV values as

$$\text{Sum CV}_c = \sum_{m \in M} CV_m(c) \tag{29}$$

Ranking of the classifiers based on their sum of CV values in ascending order are presented in Table 6. We next combine the CV scores for each classifier across all metrics to generate a global ranking. Let G be the set of global rankings for each classifier, where $|G| = 10$ (in our case). For each classifier $c \in C$, calculate the global ranking:

$$\text{Global Ranking}_c = \sum_{m \in M} CV_m(c) \tag{30}$$

Finally, rank the classifiers based on their global ranking in descending order is given in Table 7.

2) SELECTING BASE-LEVEL CLASSIFIERS WITH MONTE CARLO CROSS VALIDATION

This section is dedicated to presenting the outcomes of the rank analysis conducted on machine learning classifiers using the Monte Carlo cross-validation technique. Table 8 provides a comprehensive overview, displaying both the mean and standard deviation of each performance metric. Additionally, Supplementary Figure S5 offers a visual representation of the classifiers' rankings, arranged in ascending order based on their CV_m scores.

For a more granular examination, Table 9 furnishes the rankings derived from individual metrics, offering insights into the comparative performance of each classifier. To consolidate these findings into a holistic perspective, Table 10 consolidates the global ranking, providing a comprehensive overview of classifier performance across multiple metrics.

The global ranking, meticulously generated through both Repeated Cross-Validation (RCV) and Monte Carlo Cross-Validation (MCCV), consistently identifies Extra Trees (ET), Random Forest (RF), XGBoost (XGB), Gaussian Naive Bayes (GNB), Multi-layer Perceptron (MLP), Logistic Regression (LR), and Support Vector Machine (SVM) as the outstanding performers. Conversely, Decision Trees (DT), Linear Discriminant Analysis (LDA), and k-Nearest Neighbors (KNN) do not attain the same level of performance, as indicated by the rankings.

Consequently, the subsequent analysis in this study places exclusive emphasis on the top seven classifiers identified—Extra Trees, Random Forest, XGBoost, Gaussian Naive Bayes, Multi-layer Perceptron, Logistic Regression, and Support Vector Machine. This strategic focus ensures a detailed examination of the most promising classifiers, contributing to a more nuanced understanding of their respective strengths and capabilities in the context of the study's objectives.

3) IDENTIFICATION OF CLASSIFIER SPECIFIC TOP K FEATURES

In this series of meticulously designed experiments, our primary focus was to conduct an extensive exploration aimed

TABLE 8. Scores of performance metrics (mean value ± standard deviation) of the machine learning classification algorithms using Monte Carlo 10-fold cross validation.

Classification Algorithm	Acc	Sn (recall)	Sp	Precision	TPR	FPR	Cohen's Kappa	F1	MCC	AUC-ROC
RF	0.85 ± 0.06	0.89 ± 0.09	0.89 ± 0.00	0.84 ± 0.07	0.67 ± 0.47	0.37 ± 0.45	0.71 ± 0.12	0.86 ± 0.06	0.71 ± 0.12	0.86 ± 0.06
LR	0.81 ± 0.05	0.78 ± 0.11	0.87 ± 0.00	0.83 ± 0.07	0.57 ± 0.42	0.38 ± 0.44	0.62 ± 0.10	0.80 ± 0.07	0.63 ± 0.10	0.81 ± 0.05
LDA	0.73 ± 0.08	0.71 ± 0.10	0.71 ± 0.00	0.76 ± 0.12	0.56 ± 0.42	0.43 ± 0.42	0.46 ± 0.17	0.73 ± 0.08	0.47 ± 0.17	0.73 ± 0.08
GNB	0.85 ± 0.06	0.89 ± 0.08	0.67 ± 0.00	0.83 ± 0.08	0.61 ± 0.44	0.44 ± 0.42	0.69 ± 0.11	0.85 ± 0.05	0.70 ± 0.11	0.85 ± 0.06
ExtraTree	0.88 ± 0.06	0.91 ± 0.07	0.81 ± 0.00	0.87 ± 0.09	0.61 ± 0.44	0.40 ± 0.43	0.77 ± 0.12	0.89 ± 0.06	0.77 ± 0.12	0.89 ± 0.06
XGB	0.86 ± 0.05	0.86 ± 0.00	0.89 ± 0.00	0.86 ± 0.08	0.64 ± 0.45	0.38 ± 0.44	0.72 ± 0.11	0.86 ± 0.05	0.72 ± 0.11	0.86 ± 0.05
KNN	0.61 ± 0.07	0.23 ± 0.08	1.00 ± 0.00	1.00 ± 0.00	0.41 ± 0.43	0.33 ± 0.47	0.23 ± 0.08	0.37 ± 0.10	0.35 ± 0.07	0.62 ± 0.04
SVM	0.80 ± 0.06	0.74 ± 0.10	0.77 ± 0.00	0.85 ± 0.09	0.61 ± 0.43	0.41 ± 0.43	0.59 ± 0.11	0.79 ± 0.08	0.61 ± 0.11	0.80 ± 0.06
MLP	0.82 ± 0.06	0.82 ± 0.08	0.80 ± 0.00	0.81 ± 0.09	0.60 ± 0.43	0.40 ± 0.43	0.63 ± 0.11	0.81 ± 0.06	0.64 ± 0.11	0.82 ± 0.06
DT	0.74 ± 0.08	0.75 ± 0.08	0.57 ± 0.00	0.74 ± 0.12	0.59 ± 0.43	0.48 ± 0.41	0.48 ± 0.15	0.74 ± 0.08	0.49 ± 0.15	0.75 ± 0.07

TABLE 9. Ranking of the classifiers based on CV score of the individual performance metrics using MCCV.

Performance Metric	Ranks									
	1	2	3	4	5	6	7	8	9	10
ACC	XGB	LR	ET	RF	GNB	MLP	SVM	DT	LDA	KNN
SP	KNN	RF	LR	XGB	ET	MLP	SVM	LDA	GNB	DT
SN	XGB	ET	GNB	MLP	RF	DT	SVM	LDA	LR	KNN
PRE	RF	LR	XGB	GNB	ET	SVM	MLP	LDA	DT	KNN
TPR	RF	XGB	SVM	MLP	GNB	ET	DT	LR	LDA	KNN
FPR	DT	GNB	LDA	SVM	ET	MLP	LR	XGB	RF	KNN
K	XGB	ET	GNB	LR	RF	MLP	SVM	DT	KNN	LDA
F1	XGB	GNB	ET	RF	MLP	LR	SVM	DT	LDA	KNN
MCC	XGB	ET	GNB	LR	RF	MLP	SVM	KNN	DT	LDA
AUC-ROC	XGB	LR	KNN	ET	RF	GNB	MLP	SVM	DT	LDA

TABLE 10. Ranking of the classifier based on Global Ranking generated from MCCV.

Rank	Classifier	Combined Score
1	XGB	23
2	ET	37
3	RF	41
4	GNB	42
5	LR	47
6	MLP	57
7	SVM	63
8	DT	75
9	KNN	81
10	LDA	84

at assessing the influence of feature importance in predicting Alzheimer's disease. The investigation centered around the utilization of the top seven base-level classifiers, thoughtfully selected through the robust methodologies of Repeated Cross-Validation (RCV) and Monte Carlo Cross-Validation (MCCV).

To unravel the intricacies of feature importance, Recursive Feature Elimination (RFE) techniques were employed, specifically ranking the 337 features pertinent to classification models such as Extra Trees (ET), Random Forest (RF), XGBoost (XGB), and Logistic Regression (LR). For Gaussian Naive Bayes (GNB), Support Vector Machine (SVM), and Multi-layer Perceptron (MLP), where feature importance details are not readily available in the sklearn package of Python, the analysis was facilitated through the application of Analysis of Variance (ANOVA).

Subsequently, the examination delved into computing scores for ten distinct performance metrics within the classification task. This comprehensive evaluation spanned a range of k values (number of features), progressing systematically from 10 to 337 with increments of 10. The detailed findings and insights gleaned from this exploration are eloquently presented in Figure 5, specifically showcasing the outcomes for ET and GNB classifier models. Supplementary figures S6 to S10 complementarily enrich this narrative by encapsulating the results for classifiers RF, XGB, LR, SVM, and MLP.

The culmination of this investigation materializes in Table 11, a consolidated presentation of scores at optimal k values for each classifier across the ten performance metrics. Notably, a significant observation emerged: at k = 200, the majority of metric scores reached their zenith for the ET classifiers. This pivotal insight prompted the judicious selection of the top 200 features for the subsequent retraining of the ET classifiers. Analogously, optimal k values were identified for other classifiers: k = 250 for RF, k = 50 for LR, k = 60 for XGB, k = 150 for SVM, k = 320 for GNB, and k = 290 for MLP. Figure 6 shows the distribution of the top k features across all the tasks for the Extra Trees (ET).

Furthermore, Supplementary Tables "feature_task_mapping_ ET.csv", "feature_task_mapping_RF.csv", "feature_task_ma pping_LR.csv", "feature_task_mapping_XGB.csv", "feature_task_mapping_SVM.csv", "feature_task_mapping_GN B.csv" and "feature_task_mapping_MLP.csv" have been thoughtfully curated to provide classifier-specific details on the top k features, ensuring transparency and facilitating a deeper understanding of the feature selection process. This nuanced approach of selecting optimal features across all tasks underscores the importance

TABLE 11. Scores at the best k values for the base-level classifiers.

Metrics/ Classifies	ET	RF	LR	XGB	SVM	GNB	MLP
	best k values (score)						
Acc	200 (0.94)	250 (0.94)	50 (0.91)	60 (0.88)	150 (0.91)	320 (0.94)	290 (0.91)
Sn	200 (0.95)	30 (0.95)	50 (0.90)	60 (0.90)	10 (0.90)	320 (1.0)	10 (0.90)
Sp	10 (0.93)	250 (0.93)	30 (0.93)	30 (1.0)	110 (0.93)	110 (0.86)	100 (0.93)
Pre	200 (0.95)	250 (0.95)	50 (0.94)	30 (1.0)	150 (0.94)	320 (0.90)	290 (0.94)
TPR	200 (0.95)	30 (0.95)	50 (0.90)	60 (0.90)	10 (0.90)	320 (1.0)	10 (0.90)
FPR	10 (0.06)	250 (0.06)	30 (0.06)	30 (0.0)	110 (0.06)	110 (0.13)	100 (0.06)
Kappa	200 (0.88)	250 (0.88)	50 (0.82)	60 (0.76)	150 (0.82)	320 (0.88)	290 (0.82)
F1-score	200 (0.95)	250 (0.95)	50 (0.92)	60 (0.90)	150 (0.92)	320 (0.95)	290 (0.92)
MCC	200 (0.88)	250 (0.88)	50 (0.82)	60 (0.76)	150 (0.82)	320 (0.88)	290 (0.82)
AUC-ROC	150 (0.98)	220 (0.97)	30 (0.94)	60 (0.91)	260 (0.97)	260 (0.95)	290 (0.97)

TABLE 12. The proposed model with different combination of base level classifiers.

Methods/ Approaches	Acc	Sn	Sp	Pre	TPR	FPR	Cohen K	F1-score	MCC	AUC-ROC
Proposed Model using RF + ET and Stacking	97.14	95.00	100.00	100	95.00	0.00	94.21	97.44	94.37	97.50
Proposed Model using RF + ET + XGB and Stacking	97.14	95.00	100.00	100	95.00	0.00	94.21	97.44	94.37	97.50
Proposed Model using RF + ET + XGB + + MLP and Stacking	97.14	95.00	100.00	100	95.00	0.00	94.21	97.44	94.37	97.50
Proposed Model using RF + ET + XGB + + MLP + SVM and Stacking	97.14	95.00	100.00	100	95.00	0.00	94.21	97.44	94.37	97.50
Proposed Model using RF + ET + XGB + + MLP + SVM + GNB and Stacking	100.00	100.00	100.00	100.00	100.00	0.00	100.00	100.00	100.00	100.00
Proposed Model using Majority Voting	94.29	95.00	93.33	95.00	95.00	0.06	88.33	95.00	88.33	94.17

TABLE 13. Comparative performance with the proposed model.

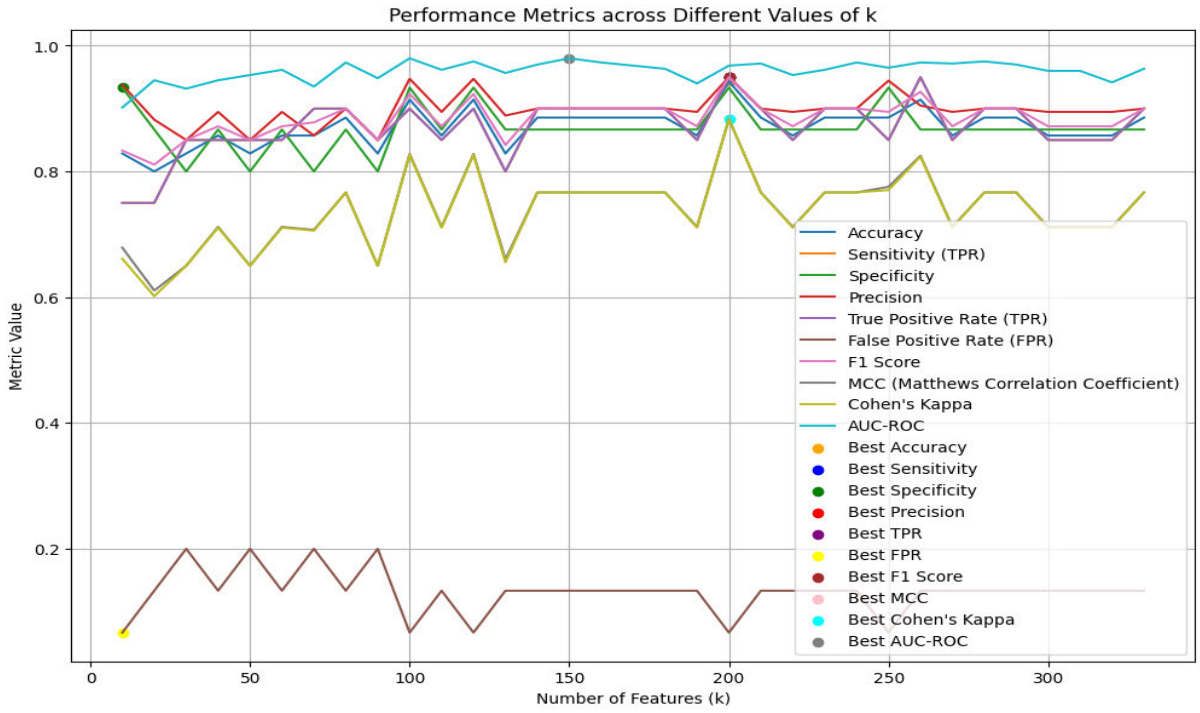
Model	**Acc (%)**	**Sn (%)**	**Sp (%)**	**Precision (%)**	**AUC-ROC (%)**
De Gregorio et al [35]	91	83	100	-	-
Cilia et al [17]	94.28	88.24	100	-	-
Parziale et al [38]	97.12	94.23	100	-	-
Subha et al [41]	90	92	-	88	90
Gattulli et al [37]	88	90	86	-	-
Onder et al [39]	85	-	-	-	-
Hakan et al [40]	97.14	-	90	95	-
Erdogmus et al [36]	90.4	-	-	-	-
Proposed Model	97.14	95	100	95	97.50

of tailoring base classifiers to specific tasks, as failure to do so, as observed in Cilia et al [17], may inadvertently lead to the exclusion of essential features critical for accurate predictive modeling.

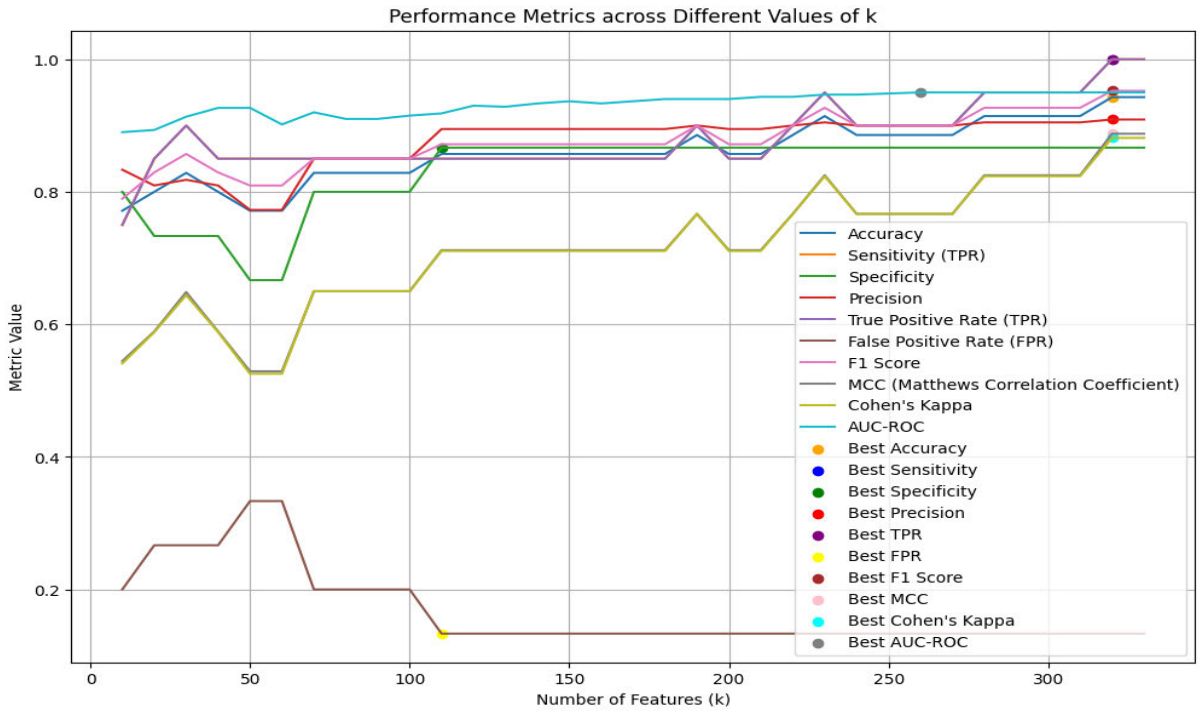
4) ENSEMBLE CLASSIFIERS

The final phase of our study involved the meticulous retraining of the top seven base-level classifiers, each using its specific subset of top-k features obtained through our rigorous feature importance analysis. To capitalize on the collective strength of these refined classifiers, we strategically implemented the stacking ensemble technique, which intelligently combines their individual predictions.

Further, we experimentally observe the influence of base level classifiers on the performance of the stacking ensemble. The results in Table 12 depict that with most combinations of base level classifiers, the stacking ensemble model achieves an accuracy of 97.14%. Notably, the inclusion of the Gaussian Naive Bayes (GNB) classifier results in a remarkable boost in the ensemble's performance, achieving a perfect accuracy of 100% and demonstrating improvements in other performance



(a)



(b)

FIGURE 5. Scores of performance metrics vs. number of top-k features (a) ET (b) GNB. Top k features for ET are selected using Recursive Feature Elimination(RFE), while ANOVA is employed for GNB.

metrics. However, the use of the GNB classifier, although straightforward and efficient, is constrained by several limitations. It assumes that features adhere to a Gaussian distribution and are independent of each other given the class label, assumptions that may not always be met in real-world scenarios. We thus used the stacking ensemble model build

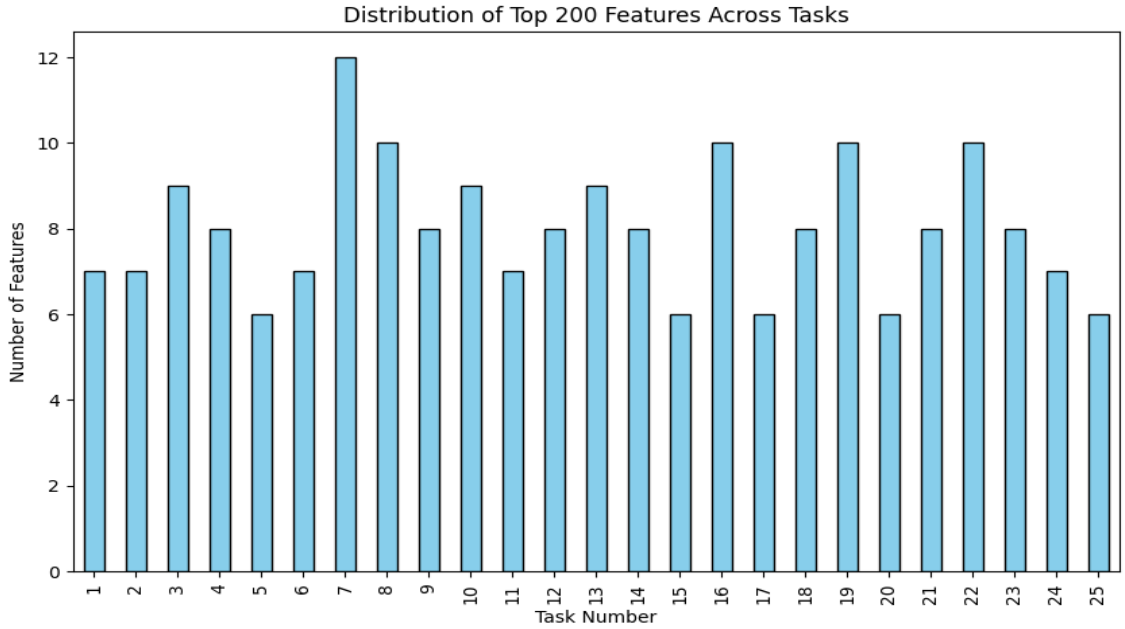


FIGURE 6. Top features across the Tasks for Extra Tree Classifier selected through Recursive Feature Elimination (RFE).

upon the Random Forest (RF) and Extra Tree Classifier (ET) as the candidate model. In parallel, we also harnessed the power of the majority voting technique to harmonize the predictions generated by the diverse base-level classifiers. The culmination of these concerted efforts is eloquently presented in the conclusive Table 12 (last row), encapsulating the amalgamated and refined predictions derived from our ensemble methodology. A juxtaposition of the outcomes with those presented in Table 12 indicates that the proposed ensemble models have successfully augmented the predictive performance.

5) COMPARATIVE PERFORMANCE ANALYSIS WITH OTHER EXISTING MODELS

Additionally, we conducted a comparative analysis with previously reported approaches focused on predicting Alzheimer's disease using the DARWIN dataset. The outcomes from these studies are detailed in Table 13, offering a side-by-side comparison with the results obtained from our proposed models. It's worth noting that the CPU time for both building and testing the ensemble model was approximately 120 seconds, indicating reasonable computational efficiency. Additionally, the memory requirement was found to be around 25 megabytes, underscoring efficient memory utilization for the task at hand. Significantly, the application of the stacking ensemble technique yielded a noteworthy enhancement in overall predictive performance. This improvement underscores the effectiveness of synergizing the strengths of diverse classifiers through a sophisticated ensemble approach, ultimately contributing to a more robust and accurate prediction of Alzheimer's disease in our study. The success observed in the final results affirms the validity of our comprehensive methodology and highlights the potential for advanced ensemble techniques in enhancing the predictive capabilities of machine learning models in medical research and diagnostic applications.

IV. CONCLUSION AND FUTURE WORK

This research successfully developed a robust and accurate predictor model for Alzheimer's disease (AD) based on handwriting analysis using machine learning techniques. The proposed ensemble model, built upon meticulously chosen base classifiers and optimized through rigorous feature selection methods, achieved impressive performance metrics achieving 97.14% accuracy, 95% sensitivity, 100% specificity, 100% Precision, 97.44% F1-score, 94.37% Matthews Correlation Coefficient (MCC), Cohen Kappa 94.21% and 97.5% Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These findings demonstrate the immense potential of machine learning in facilitating early and accurate diagnosis of AD through a non-invasive and readily accessible method like handwriting analysis.

It has significant implications for clinical practice, allowing for earlier intervention, improved disease management, and ultimately, a higher quality of life for individuals with AD. Future research could explore incorporating additional data modalities and leveraging more advanced machine learning algorithms to further enhance the accuracy and robustness of the predictive model. Additionally, investigating the feasibility of using mobile technology for real-time monitoring of handwriting patterns could open doors for personalized intervention strategies and early detection of AD

progression. By advancing the field of AD diagnosis through innovative and accessible methods, we can pave the way for improved patient care and a brighter future for individuals living with this challenging condition.

DATA AVAILABILITY

The dataset is openly accessible at https://archive.ics.uci.edu/dataset/732/darwin.

CODE AVAILABILITY

Codes for implementing, training and testing, are available at https://github.com/shafiqulrehman/AlzheimerPredictionML Code

ACKNOWLEDGMENT

The authors would like to acknowledge and express their appreciation for the support extended by Google, providing free access to the Colab runtime environment. This assistance has significantly contributed to the efficiency of their work. Furthermore, the authors extend their heartfelt gratitude to the UCI Machine Learning Repository for making the DARWIN dataset openly accessible. Without this invaluable resource, the progress and successful completion of this endeavor would have been unattainable.

REFERENCES

[1] *Dementia*. Accessed: Nov. 13, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[2] M. J. Armstrong, I. Litvan, A. E. Lang, T. H. Bak, K. P. Bhatia, B. Borroni, A. L. Boxer, D. W. Dickson, M. Grossman, M. Hallett, and K. A. Josephs, "Criteria for the diagnosis of corticobasal degeneration," *Neurology*, vol. 80, no. 5, pp. 496–503, 2013.

[3] R.-E. Precup, T.-A. Teban, A. Albu, A.-B. Borlea, I. A. Zamfirache, and E. M. Petriu, "Evolving fuzzy models for prosthetic hand myoelectric-based control," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 7, pp. 4625–4636, Jul. 2020.

[4] G. Vessio, "Dynamic handwriting analysis for neurodegenerative disease assessment: A literary review," *Appl. Sci.*, vol. 9, no. 21, p. 4666, Nov. 2019.

[5] C. De Stefano, F. Fontanella, D. Impedovo, G. Pirlo, and A. S. di Freca, "Handwriting analysis to support neurodegenerative diseases diagnosis: A review," *Pattern Recognit. Lett.*, vol. 121, pp. 37–45, Apr. 2019.

[6] P. Singh and H. Yadav, "Influence of neurodegenerative diseases on handwriting," *Forensic Res. Criminol. Int. J.*, vol. 9, no. 3, pp. 110–114, 2021.

[7] D. Impedovo, G. Pirlo, G. Vessio, and M. T. Angelillo, "A handwriting-based protocol for assessing neurodegenerative dementia," *Cogn. Comput.*, vol. 11, pp. 576–586, Aug. 2019.

[8] D. Impedovo, G. Pirlo, and G. Vessio, "Dynamic handwriting analysis for supporting earlier Parkinson's disease diagnosis," *Information*, vol. 9, no. 10, p. 247, Oct. 2018.

[9] F. Cavallo, A. Moschetti, D. Esposito, C. Maremmani, and E. Rovini, "Upper limb motor pre-clinical assessment in Parkinson's disease using machine learning," *Parkinsonism Rel. Disorders*, vol. 63, pp. 111–116, Jun. 2019.

[10] M. A. Myszczynska, P. N. Ojamies, A. M. B. Lacoste, D. Neil, A. Saffari, R. Mead, G. M. Hautbergue, J. D. Holbrook, and L. Ferraiuolo, "Applications of machine learning to diagnosis and treatment of neurodegenerative diseases," *Nature Rev. Neurol.*, vol. 16, no. 8, pp. 440–456, Aug. 2020.

[11] M. Belić, V. Bobić, M. Badža, N. Šolaja, M. Durić-Jovičić, and V. S. Kostić, "Artificial intelligence for assisting diagnostics and assessment of Parkinson's disease—A review," *Clin. Neurol. Neurosurgery*, vol. 184, Sep. 2019, Art. no. 105442.

[12] I. Scott, S. Carter, and E. Coiera, "Clinician checklist for assessing suitability of machine learning applications in healthcare," *BMJ Health Care Informat.*, vol. 28, no. 1, Feb. 2021, Art. no. e100251.

[13] Y. A. Nanehkaran, D. Zhang, S. Salimi, J. Chen, Y. Tian, and N. Al-Nabhan, "Analysis and comparison of machine learning classifiers and deep neural networks techniques for recognition of Farsi handwritten digits," *J. Supercomput.*, vol. 77, no. 4, pp. 3193–3222, Apr. 2021.

[14] Y. A. Nanehkaran, J. Chen, S. Salimi, and D. Zhang, "A pragmatic convolutional bagging ensemble learning for recognition of Farsi handwritten digits," *J. Supercomput.*, vol. 77, no. 11, pp. 13474–13493, Nov. 2021.

[15] T. M. Ghanim, M. I. Khalil, and H. M. Abbas, "Comparative study on deep convolution neural networks DCNN-based offline Arabic handwriting recognition," *IEEE Access*, vol. 8, pp. 95465–95482, 2020.

[16] J. Memon, M. Sami, R. A. Khan, and M. Uddin, "Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR)," *IEEE Access*, vol. 8, pp. 142642–142668, 2020.

[17] N. D. Cilia, G. De Gregorio, C. De Stefano, F. Fontanella, A. Marcelli, and A. Parziale, "Diagnosing Alzheimer's disease from on-line handwriting: A novel dataset and performance benchmarking," *Eng. Appl. Artif. Intell.*, vol. 111, May 2022, Art. no. 104822.

[18] N. Ranjan, D. Umesh, V. Dongare, K. Chavan, and Y. Kuwar, "Diagnosis of Parkinson disease using handwriting analysis," *Int. J. Comput. Appl.*, vol. 184, no. 1, pp. 13–16, Mar. 2022.

[19] S. Qasim Abbas, L. Chi, and Y.-P.-P. Chen, "Transformed domain convolutional neural network for Alzheimer's disease diagnosis using structural MRI," *Pattern Recognit.*, vol. 133, Jan. 2023, Art. no. 109031.

[20] S. Ul Rehman, N. Tarek, C. Magdy, M. Kamel, M. Abdelhalim, A. Melek, L. N. Mahmoud, and I. Sadek, "AI-based tool for early detection of Alzheimer's disease," *Heliyon*, vol. 10, no. 8, Apr. 2024, Art. no. e29375.

[21] J. Silva, B. C. Bispo, and P. M. Rodrigues, "Structural MRI texture analysis for detecting Alzheimer's disease," *J. Med. Biol. Eng.*, vol. 43, no. 3, pp. 227–238, 2023.

[22] V. P. S. Rallabandi and K. Seetharaman, "Deep learning-based classification of healthy aging controls, mild cognitive impairment and Alzheimer's disease using fusion of MRI-PET imaging," *Biomed. Signal Process. Control*, vol. 80, Feb. 2023, Art. no. 104312.

[23] H. B. Baydargil, J. Park, and I. F. Ince, "Anomaly-based Alzheimer's disease detection using entropy-based probability positron emission tomography images," *ETRI J.*, pp. 1–13, Mar. 2024.

[24] H. Ahmed, H. Soliman, and M. Elmogy, "Early detection of Alzheimer's disease using single nucleotide polymorphisms analysis based on gradient boosting tree," *Comput. Biol. Med.*, vol. 146, Jul. 2022, Art. no. 105622.

[25] W. Yin, T. Yang, G. Wan, and X. Zhou, "Identification of image genetic biomarkers of Alzheimer's disease by orthogonal structured sparse canonical correlation analysis based on a diagnostic information fusion," *Math. Biosciences Eng.*, vol. 20, no. 9, pp. 16648–16662, 2023.

[26] A. S. Alatrany, W. Khan, A. J. Hussain, J. Mustafina, and D. Al-Jumeily, "Transfer learning for classification of Alzheimer's disease based on genome wide data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 20, no. 5, pp. 2700–2711, May 2023.

[27] B. Jiao, R. Li, H. Zhou, K. Qing, H. Liu, H. Pan, Y. Lei, W. Fu, X. Wang, X. Xiao, X. Liu, Q. Yang, X. Liao, Y. Zhou, L. Fang, Y. Dong, Y. Yang, H. Jiang, S. Huang, and L. Shen, "Neural biomarker diagnosis and prediction to mild cognitive impairment and Alzheimer's disease using EEG technology," *Alzheimer's Res. Therapy*, vol. 15, no. 1, pp. 1–14, 2023.

[28] Q. Hou, Y. Guan, X. Liu, M. Xiao, and Y. Lü, "Development and validation of a risk model for cognitive impairment in the older Chinese inpatients: An analysis based on a 5-year database," *J. Clin. Neurosci.*, vol. 104, pp. 29–33, Oct. 2022.

[29] S. Qiu et al., "Multimodal deep learning for Alzheimer's disease dementia assessment," *Nature Commun.*, vol. 13, no. 1, p. 3404, 2022.

[30] C. Pozna and R. E. Precup, "Applications of signatures to expert systems modelling," *Acta Polytechnica Hungarica*, vol. 11, no. 2, pp. 21–39, 2014.

[31] G. Pirlo, M. Diaz, M. A. Ferrer, D. Impedovo, F. Occhionero, and U. Zurlo, "Early diagnosis of neurodegenerative diseases by handwritten signature analysis," in *Proc. Int. Workshops New Trends Image Anal. Process. (ICIAP)*, Genoa, Italy. Fisciano, Italy: Springer, Sep. 2015, pp. 290–297.

[32] N. D. Cilia, C. D. Stefano, F. Fontanella, and A. S. Di Freca, "An experimental protocol to support cognitive impairment diagnosis by using handwriting analysis," *Proc. Comput. Sci.*, vol. 141, pp. 466–471, Jan. 2018.

[33] R. Senatore and A. Marcelli, "A paradigm for emulating the early learning stage of handwriting: Performance comparison between healthy controls and Parkinson's disease patients in drawing loop shapes," *Human Movement Sci.*, vol. 65, pp. 89–101, Jun. 2019.

[34] G. De Gregorio, D. Desiato, A. Marcelli, and G. Polese, "A multi classifier approach for supporting Alzheimer's diagnosis based on handwriting analysis," in *Proc. ICPR Int. Workshops Challenges Pattern Recognit.* Springer, Jan. 2021, pp. 559–574.
[35] P. Erdogmus and A. T. Kabakus, "The promise of convolutional neural networks for the early diagnosis of the Alzheimer's disease," *Eng. Appl. Artif. Intell.*, vol. 123, Aug. 2023, Art. no. 106254.
[36] V. Gattulli, D. Impedovo, and G. P. G. Semeraro, "Handwriting task-selection based on the analysis of patterns in classification results on Alzheimer dataset," in *Proc. IEEESDS*, Jun. 2023, pp. 1–12.
[37] A. Parziale, A. D. Cioppa, and A. Marcelli, "Investigating one-class classifiers to diagnose Alzheimer's disease from handwriting," in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, May 2022, pp. 111–123.
[38] M. Önder, Ü. Şentürk, K. Polat, and D. Paulraj, "Diagnosis of Alzheimer's disease using boosting classification algorithms," in *Proc. Int. Conf. Res. Methodol. Knowl. Manage., Artif. Intell. Telecommun. Eng. (RMKMATE)*, Nov. 2023, pp. 1–5.
[39] H. Öcal, "A novel approach to detection of Alzheimer's disease from handwriting: Triple ensemble learning model," *Gazi Univ. J. Sci. C, Design Technol.*, vol. 12, no. 1, pp. 214–223, Mar. 2024.
[40] R. Subha, B. R. Nayana, and M. Selvadass, "Hybrid machine learning model using particle swarm optimization for effectual diagnosis of Alzheimer's disease from handwriting," in *Proc. 4th Int. Conf. Circuits, Control, Commun. Comput. (I4C)*, Dec. 2022, pp. 491–495.
[41] V. Viswan, N. Shaffi, M. Mahmud, K. Subramanian, and F. Hajamohideen, "Explainable artificial intelligence in Alzheimer's disease classification: A systematic review," *Cogn. Comput.*, vol. 16, no. 1, pp. 1–44, Jan. 2024.
[42] N. Shaffi, K. Subramanian, V. Vimbi, F. Hajamohideen, A. Abdesselam, and M. Mahmud, "Performance evaluation of deep, shallow and ensemble machine learning methods for the automated classification of Alzheimer's disease," *Int. J. Neural Syst.*, pp. 1–16, Apr. 2024.

UDDALAK MITRA received the Ph.D. degree from Visva-Bharati University, Santiniketan, India. He is currently an Assistant Professor with Siliguri Institute of Technology, MAKAUT University, Kolkata, West Bengal. He is also a specialized professional in bioinformatics, computational biology, machine learning and deep learning, and applied artificial intelligence (ML/DL) in the field of agriculture and medical diagnosis. He has authored and coauthored more than 14 papers in journals, conference proceedings, and book chapters, and supervises Ph.D. students, in addition to master's and undergraduate students. He is also a regular reviewer in reputed international and peer-reviewed journals.

SHAFIQ UL REHMAN received the Ph.D. degree from Universiti Sains Malaysia (USM), Malaysia, in 2017. He was a Postdoctoral Research Fellow with Singapore University of Technology and Design (SUTD), Singapore, from 2017 to 2020. He is currently the Chairperson of the Department of Computer Science, College of Information Technology, Kingdom University, Bahrain. He is also an Assistant Professor specializing in cybersecurity, artificial intelligence, the Internet of Things (IoT), industry 4.0, and cloud/edge computing. He has authored and coauthored more than 50 papers in journals, conference proceedings, and book chapters, and supervises Ph.D., postgraduate, and undergraduate students. He is involved in various research projects related to secure machine-to-machine communication, the IoT in healthcare, industry 4.0, and emerging technologies using open-source platforms. He has experience building AI tools for healthcare, security systems for communication protocols, the IoT devices, and cloud/edge computing architecture.