Optimized Transfer Learning Based Dementia Prediction System for Rehabilitation Therapy Planning

Ping-Huan Kuo, *Associate Member, IEEE*, Chen-Ting Huang, and Ting-Chun Yao

Abstract— **Dementia is a neurodegenerative disease that causes a progressive deterioration of thinking, memory, and the ability to perform daily tasks. Other common symptoms include emotional disorders, language disorders, and reduced mobility; however, self-consciousness is unaffected. Dementia is irreversible, and medicine can only slow but not stop the degeneration. However, if dementia could be predicted, its onset may be preventable. Thus, this study proposes a revolutionary transfer-learning machine-learning model to predict dementia from magnetic resonance imaging data. In training, k-fold cross-validation and various parameter optimization algorithms were used to increase prediction accuracy. Synthetic minority oversampling was used for data augmentation. The final model achieved an accuracy of 90.7%, superior to that of competing methods on the same data set. This study's model facilitates the early diagnosis of dementia, which is key to arresting neurological deterioration from the disease, and is useful for underserved regions where many do not have access to a human physician. In the future, the proposed system can be used to plan rehabilitation therapy programs for patients.**

Index Terms— **Dementia prediction, machine learning, parameter optimization, transfer learning.**

I. INTRODUCTION

DEMENTIA is a chronic degenerative disease [1], [2], [3], [4] characterized by a progressive and irreversible decline in brain function; in particular, it induces behavioral

Ping-Huan Kuo is with the Department of Mechanical Engineering, National Chung Cheng University, Chiayi 62102, Taiwan, and also with the Advanced Institute of Manufacturing with High-Tech Innovations (AIM-HI), National Chung Cheng University, Chiayi 62102, Taiwan (e-mail: phkuo@ccu.edu.tw).

Chen-Ting Huang is with the Department of Mechanical Engineering, National Chung Cheng University, Chiayi 62102, Taiwan (e-mail: irisrose123456789@gmail.com).

Ting-Chun Yao is with the School of Medicine, College of Medicine, Taipei Medical University, Taipei City 11031, Taiwan (e-mail: jim30081212@gmail.com).

changes and impedes a patient's ability to perform activities of daily living (ADLs). Dementia affects millions of people worldwide and is becoming more prevalent as the planet's population ages. The World Alzheimer Report 2019, published by Alzheimer's Disease (AD) International, estimated that more than 50 million people live with dementia globally. They estimated that this number will increase to 152 million by 2050, equivalent to one person developing dementia every 3 s. Nonetheless, rapid and timely diagnosis can slow this decline in brain function. Manual tools for predicting dementia are inaccurate [2], [5], [6], complex, and require cognitive tests to be administered over a long time. Therefore, previous studies have formulated machine-learning tools [7], [8]. based on the k-nearest neighbor, decision tree, support vector machine (SVM), and extreme gradient boosting (XGBoost) approaches; these tools have been extensively used for rapid and timely diagnosis and clinical decision-making.

One study [9] used an algorithm to distinguish healthy participants from participants with dementia on the basis of behavioral data; in a sequence prediction task, participants with dementia had significantly lower peak accuracy scores (11%) than healthy patients. Sequential pattern discovery using equivalence classes was employed to identify various parameters for early-stage dementia diagnosis. The algorithm could detect early dementia symptoms without the need for expensive clinical procedures. In contrast to the aforementioned study, [10] formulated a method that uses language samples instead. They considered speech and language impairments, which are common in several neurodegenerative diseases, in their cognitive impairment analysis to achieve early diagnosis and identify the onset of cognitive decline. They further introduced several original lexical and syntactic features in addition to a previously established lexical syntax to train machine-learning classifiers to identify the etiologies of AD, mild cognitive impairment (MCI) and possible AD (PoAD). A decline in linguistic function is associated with neurodegenerative diseases and cognitive decline, and the statistical analysis of lexicosyntactic biomarkers may facilitate the early diagnosis of these diseases. Dementia is closely related to cognitive impairment, but cognitive impairment does not necessarily lead to dementia. According to a report by the Chang Gung Dementia Center, MCI is a transitional period during

which the cognitive function of the patient differs from that of a normal older adult. The probability of this MCI progressing to dementia is approximately 10%–15%, far greater than 1%–2% for a group of individuals without MCI. Electroencephalography (EEG) signals obtained during cognitive tests have also been subject to iterative filtering decomposition for dementia prediction [2]. Continuous EEGs were recorded in two resting states (i.e., eyes open and closed) and two cognitive states (i.e., finger-tapping test and continuous performance test). The EEG signals were decomposed using iterative filtering, and four key EEG features were used for multiclass classification. The method was effective for the early diagnosis and prediction of dementia and was superior to decision tree, k-nearest neighbor, SVM, and ensemble classifiers. Similarly, [11] proposed a method for early prediction of dementia by using an innovative travel pattern classification. Environmental passive sensor signals were employed to sense the movements of the inhabitants of a space. The system segmented the movements into travel episodes and classified them using a recurrent neural network. The recurrent neural network was selected because it can process raw movement data directly and does not require domain-specific knowledge for feature engineering. Finally, imbalance in the data with respect to travel pattern classes was handled using the focal loss, and the discriminative ability of the deep-learning features was enhanced using a center loss function. Multiple experiments were performed on real-life datasets to verify the system's accuracy. Another study [6] used the XGBoost algorithm to predict dementia risk. The XGBoost-based dementia risk prediction model was constructed using variables extracted from quantitative data on dementia, and its hyperparameters were optimized. This method generates top-N groups by extracting the most important variables. Hyperparameter optimization was performed in accordance with the features of the data for each top-N group. The performance of the XGBoost-based model in determining dementia risk was evaluated using the group with the best performance.

This study employed transfer learning and parameter optimization algorithms to produce a dementia prediction model. In the transfer learning framework of this model, multiple weak classifiers were combined into a strong classifier to reduce training time and expedite data aggregation. This framework was integrated with parameter optimization algorithms to improve model accuracy without the need to adjust relevant parameters manually. Other models, namely multilayer perceptron (MLP) [12], random forest [13], support vector classification (SVC) [14], AdaBoost [15], and XGBoost [16], were also used for model training and were used in evaluations of the proposed transfer-learning model. The results of this study were also compared with the prediction results of [6] and [17], which were based on the same dataset. In these comparisons, the accuracy of the proposed model was higher than that of the other models. In addition, various parameter optimization algorithms were applied to improve the accuracy of the final model. This study's model facilitates the early diagnosis of dementia, which is key to arresting neurological deterioration from the disease, and is useful for underserved regions where many do not have access to a human physician. Section II of this study introduces the system's architecture, including the content and sorting method of the data sets and correlations among internal parameters of the data sets. Section III explains the model training, including the principle underlying the model algorithm and the parameter settings for each model. Section IV presents the prediction results and compares them with those of other models. Section V provides a discussion of the results. Finally, Section VI organizes the results presented in Section IV and includes the conclusions of this study.

II. SYSTEM ARCHITECTURE

The data sets employed in this study were obtained from the Open Access Series of Imaging Studies (OASIS), a series of neuroimaging data sets that are publicly available for research and analysis [18]. The data sets contain numerical brain magnetic resonance imaging (MRI) data from right-handed individuals with and without dementia and aged 60–96 years. The sample comprised 150 individuals (both sexes) who underwent two or more MRI scans 1 year apart for a total of 373 MRI scans. The variables in the data set are presented in Table I [6], these are number of MRI scans, time interval between two or more MRI scans, sex, age, years of education, socioeconomic status (SES), mini-mental state examination (MMSE) score, clinical dementia rating (CDR), estimated total intracranial volume (eTIV), normalized whole brain volume (nWBV), and Atlas scaling factor (ASF). SES was scored from 1 (low) to 5 (high). The Mini-Mental State Examination (MMSE) score [19] were used to indicate cognitive ability and dementia severity and range from 0 (highest risk of dementia) to 30. CDR is an evaluation of six items, namely memory, orientation, judgment, and problem solving, community affairs, home and hobbies, and personal care; memory is the main evaluated item. The eTIV, nWBV, and ASF values were extracted from the MRI data. eTIV is the estimated total intracranial volume, nWBV is the normalized whole brain volume, and ASF is the Atlas scaling factor.

Preprocessing was performed prior to model training to remove unnecessary variables and data. All individuals were right-handed; hence, this variable was removed. The numerical patient identifier was also removed. Finally, some individuals had missing data, and their data were removed from the data set. A correlation analysis was conducted on the remaining variables.

Fig. 1 reveals that the remaining variables were somewhat correlated; hence, they were retained.

The dementia prediction method of this study proceeds per the flow chart in Fig. 2. First, an OASIS brain MRI data set was preprocessed by deleting irrelevant or missing data, quantizing the data, and normalizing the data. A linear transformation was applied to normalize the data for each variable to range from 0 to 1. This transformation process not only retained the original data sequence but also facilitated interdimensional comparisons and improved classification accuracy. Model training was performed using k-fold cross-validation in which the data are divided into k equal parts. The model is
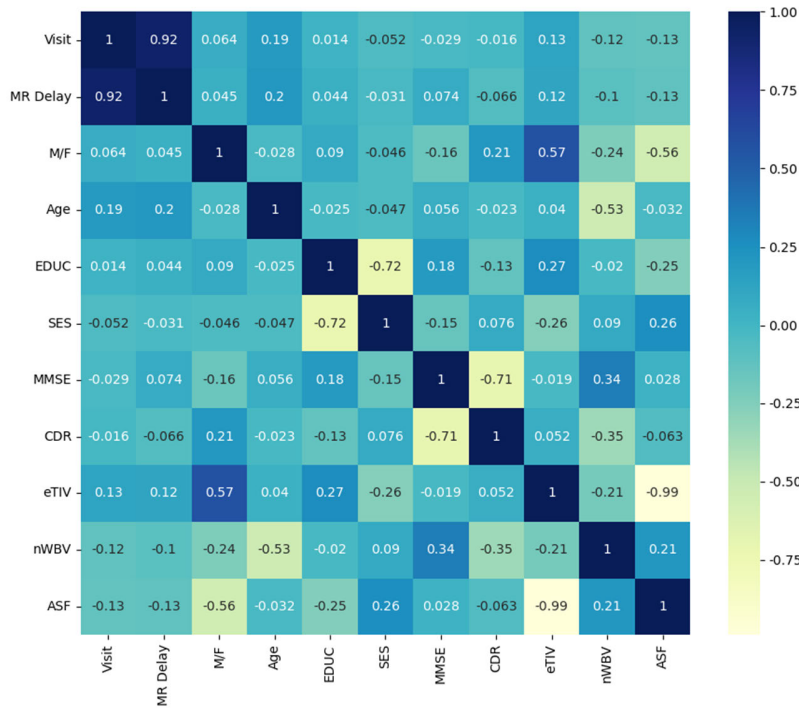
Fig. 1. Correlations among parameters in the data sets.

TABLE I
DATA PARAMETER

| Parameter	Description
Visit	Number of MRI scans
MR Delay	Time interval between two or more MRI scans
M/F	Sex
Age	Age
EDUC	Years of education
SES	Socioeconomic status; highest status: 1; lowest status: 5
MMSE	Mini-Mental State Examination score
CDR	Clinical dementia rating: 0, normal; 0.5, very mild dementia; 1, mild dementia; 2, moderately severe dementia; and 3, severe dementia
eTIV	Estimated total intracranial volume (from MRI)
nWBV	Normalized whole brain volume (determined using MRI data)
ASF	Atlas scaling factor
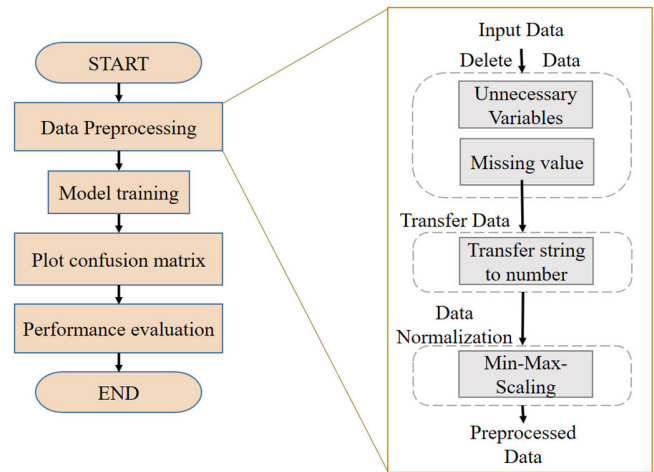
Fig. 2. Flowchart of dementia predictive analysis.

Fig. 3. Confusion matrix.

trained k times on the training set comprising k − 1 data sets, and the remaining part is the test set. Prediction consistency is improved if all data points in the test set are used once. The results of various models were compared and analyzed using confusion matrices. Moreover, the results of the proposed model were compared with those of other studies that used the same data sets.

III. METHOD

This study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Human Research Ethics Committee, National Chung Cheng University (Application number: CCUREC111090601). A confusion matrix was used to visualize the prediction results. A two-class confusion matrix is presented in Fig. 3. The columns and rows of the matrix represent predicted and ground-truth class instances, respectively, and indicates whether the model makes erroneous predictions.

For two classes, the accuracy is calculated as follows:

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

Precision is the probability that the model predicts true given that the ground truth is true; recall is the probability that the ground truth is true given that the model predicts true. The equations for calculating precision and recall are as follows:

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

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

The F1 score combines the precision and recall; it ranges from 0 to 1 with 1 representing perfect output and 0 representing the worst possible output. It is calculated as follows:

$$F1\_score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

However, this study had three classification targets (i.e., Class 0 for nondementia, Class 1 for dementia, and Class 2 for conversion to dementia); hence, precision, recall, and F1-Score could not be calculated; only the accuracy could be determined. Moreover, these classes were of unequal size; hence, the data were imbalanced.

This study employed k-fold cross validation to prevent errors in the training results caused by specific data. In this method, the data are divided into k equal parts, and one part is employed as the test data each time, with the remaining $k-1$ parts used as the training data. Training is performed k times until each set of data has been tested, as illustrated in Fig 4(a). The results of each test were represented by a confusion matrix. In this study, stratified k-fold cross-validation was used to deal with data imbalance; each of the k subsets had the same proportion of data from each class as the overall data set Fig 4(b). This method reduces prediction error due to data imbalance.

To further minimize the effect of data imbalance on model training, the synthetic minority oversampling technique (SMOTE) was employed to generate new minority samples which were included in the training data to improve predictions for imbalanced classes [20]. Fig. 5 illustrates how synthetic samples are generated in SMOTE. The gray and blue circles in Fig. 5 compose the majority and minority classes, respectively. One of the minority circles is selected, and the k points nearest this circle are identified; one of these is then selected at random, and a new sample is then generated between these two points. To generate new categorical data, averaging could not be used (e.g., the feature data of 0 and 1 cannot be 0.5); instead, the SMOTE–nominal continuous method was employed in which the highest-frequency data value among the adjacent points is used for the synthetic sample. SMOTE was used to expand all classes except the largest class for the features of dementia type, sex, and SES. The number of neighbors k was set to 5, which means the sample's neighborhood is defined by the closest five neighbors to generate synthetic samples.

Various parameter optimization algorithms, namely gray wolf optimization (GWO), the genetic algorithm (GA),
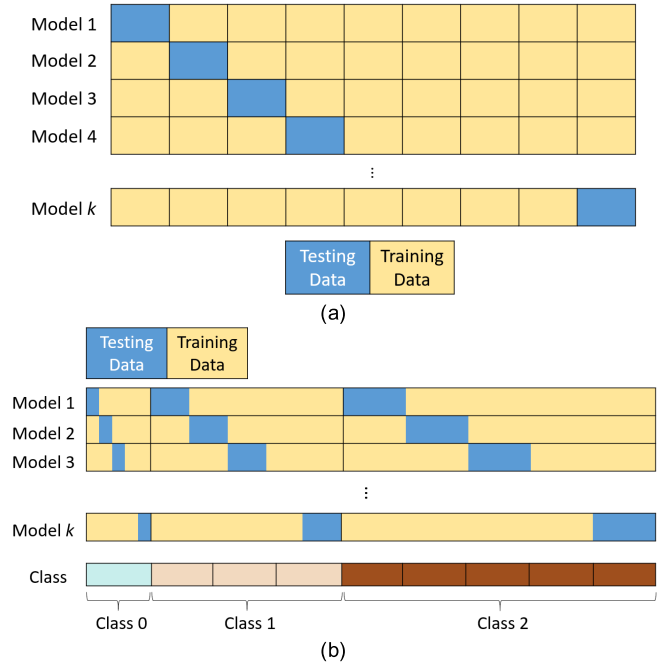

Fig. 4. The types of cross-validation (a) k-Fold Cross-Validation. (b) Stratified k-Fold Cross-Validation.

monarch butterfly optimization (MBO), and particle swarm optimization (PSO), were used to optimize model hyperparameters, and the optimized models were compared.

GWO mimics the leadership hierarchy and hunting mechanism of gray wolves in nature [21]. The leadership hierarchy is represented by α, β, δ, and ω, and the hunting mechanism comprises three steps: encircling, chasing, and attacking prey. In the algorithm, α is the optimal solution, β and δ are the second and third most optimal solutions in sequence, and ω is guided by α, β, and δ. The governing equations are as follows:

$$\overrightarrow{D} = \left| \overrightarrow{C} \cdot \vec{X}_p(t) - \vec{X}(t) \right| \quad (5)$$

$$\vec{X}(t+1) = \vec{X}_p(t) - \vec{A} \cdot \overrightarrow{D} \quad (6)$$

$$\vec{A} = 2\overrightarrow{a} \cdot \overrightarrow{r_1} - \vec{a} \quad (7)$$

$$\vec{C} = 2 \cdot \overrightarrow{r_2} \quad (8)$$

where $\overrightarrow{D}$ is the distance between an individual wolf and the prey; $\vec{X}(t+1)$ is the updated position of the wolf; t is the current number of iterations; $\vec{X}_p$ and $\vec{X}$ are the position vectors of the prey and wolf, respectively; $\vec{A}$ and $\vec{C}$ are coefficient vectors and $\vec{r}_1$ and $\vec{r}_2$ are random vectors in [0, 1]. The positions of the wolves are updated on the basis of the distance from the prey. To promote exploration (searching) and exploitation (attacking), $\vec{a}$ decreases linearly from 2 to 0 over iterations. If $|A| \leq 1$, the wolves will be closer to the prey in the next position, whereas $|A| > 1$ indicates that the wolves are moving away from the prey [22]. When the maximum number of iterations is reached, the optimal solution is output, and the algorithm terminates. A flowchart of the algorithm is displayed in Fig. 6.

The GA was developed on the basis of biological concepts, and it uses selection, crossover, and mutation operations to
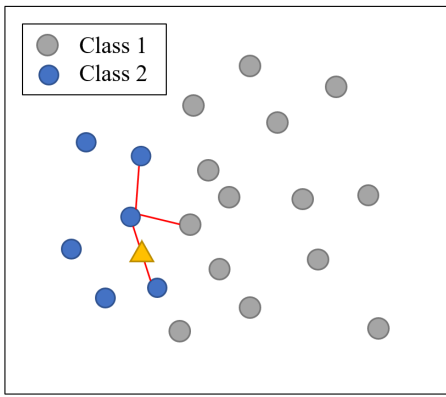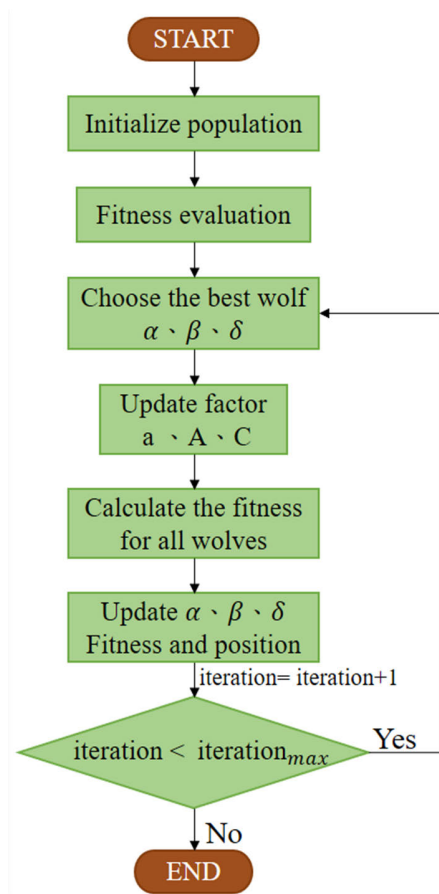
Fig. 5. SMOTE sample generation.
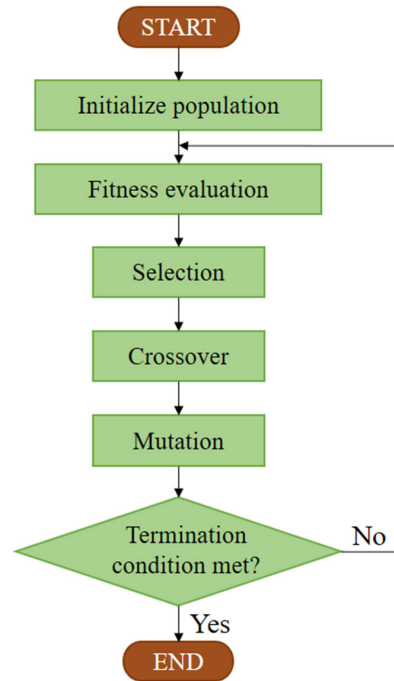

Fig. 6. GWO flowchart.

Fig. 7. GA flowchart.

MBO is idealized from the migration behavior of monarch butterflies [25], [26] in accordance with the following rules: (1) The entire monarch butterfly population comprises butterflies in Subpopulations 1 and 2. (2) Each offspring is generated by the monarch butterfly migration operator from these two subpopulations. (3) The number of butterflies is unchanged during the optimization process. (4) The fittest individual butterfly cannot be updated by any operator. The MBO algorithm mainly comprises a migration operator and adjustment operator.

For the monarch butterflies in Subpopulation 1, the migration operator is represented by the following equations:

$$x_{i,k}^{t+1} = \begin{cases} x_{r_1,k}^t & r \le p \\ x_{r_2,k}^t & r > p \end{cases} \tag{9}$$

$$\text{r} = R_{\text{rand}} \, T_{\text{peri}} \tag{10}$$

where p is the proportion of monarch butterflies in Subpopulation 1, and $x_{r_1,k}^t$ represents the new position of individual r_1 in dimension k in and iteration t and is only established when $r \le p$. In addition, $\text{r} = R_{rand} * T_{peri}$, where T_{peri} represents the transitional period and R_{rand} represents a random number in [0,1]. By contrast, $x_{r_2,k}^t$ is updated on the basis of an individual r_2 randomly selected from Subpopulation 2 when $r > p$. Individuals r_1 and r_2 are selected from Subpopulations 1 and 2, respectively.

The adjustment operator is crucial for all individuals in Subpopulation 2. The position of each individual is updated as follows:

$$x_{j,k}^{t+1} = \begin{cases} x_{\text{best},k}^t & R_{\text{rand}\,1} \le p \\ x_{r_3,k}^t & R_{\text{rand}\,1} > p \\ x_{j,k}^{t+1} + \alpha\,(d_k - 0.5) & R_{\text{rand}\,1} > p \text{ and } R_{\text{rand}\,2} > R_{BAR} \end{cases} \tag{11}$$

identify and produce genes that are fit the environment. All three operations affect diversity; selection refers to selection of the best gene, which reduces diversity; crossover has no effect on diversity; and mutation increases diversity [23]. Starting from an initial population, all genes are evaluated using a fitness function [24], and genes are selected for crossover and mutation to generate new individuals, which may be improved individuals. The improved individuals replace the initial population, and the process is repeated until the maximum number of iterations is reached, at which point the optimal solution is output and the algorithm terminates. A flowchart of the process is presented in Fig. 7.
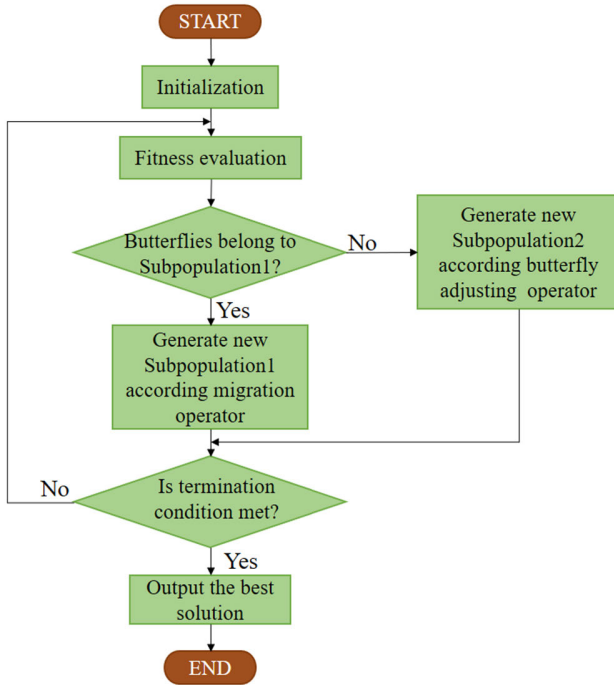
Fig. 8.   MBO flowchart.

Fig. 9.   PSO flowchart.

$$\alpha = S_{\max}/t^2 \tag{12}$$

$$d = \text{Levy}\left(x_j^t\right) \tag{13}$$

If $R_{rand1} \leq p$, $x_{j,k}^{t+1}$ is updated to $x_{best,k}^t$, where $x_{best,k}^t$ is the optimal solution of groups in dimension k after t iterations. If $R_{rand2} > p$, the position of individual r_3 is updated, where r_3 is a randomly selected individual from Subpopulation 2. If $R_{rand1} > p$ and $R_{rand2} > R_{BAR}$, the update method $x_{j,k}^{t+1} = x_{j,k}^{t+1} + \alpha (d_k - 0.5)$ is employed, where R_{BAR} is the adjustment rate and d_k is the walk step obtained by computing the Lévy flight. In the calculation of the weighting factor α, S_{max} is the maximum distance that an individual can move in one step. Migration is executed iteratively, and the operators are adjusted until the maximum number of iterations is reached, at which point the optimal solution is output and the algorithm terminates. A flowchart of the MBO process is shown in Fig. 8.

The PSO algorithm [27] proceeds as follows. In this algorithm, particles in a group are initialized with random velocities. Each particle then searches in the problem space to improve its position and is updated to this position. The optimal solution is the global optimal position. PSO is an optimization method in which improvements are made through continual iteration. The velocity of a particle (i.e., the position) is calculated using the following equations:

$$X_i^{k+1} = X_i^k + V_i^{k+1} \tag{14}$$

$$V_i^{k+1} = \omega V_i^k + c_1 r_1 \times \left(\text{pbest } t_i^k - X_i^k \right)$$
$$+ c_2 r_2 \times \left(\text{gbest }_i^k - X_i^k \right) \tag{15}$$

where ω is the weight, c_1 and c_2 are acceleration constants, r_1 and r_2 are random values in the interval [0,1], V_i^{k+1} and X_i^{k+1}
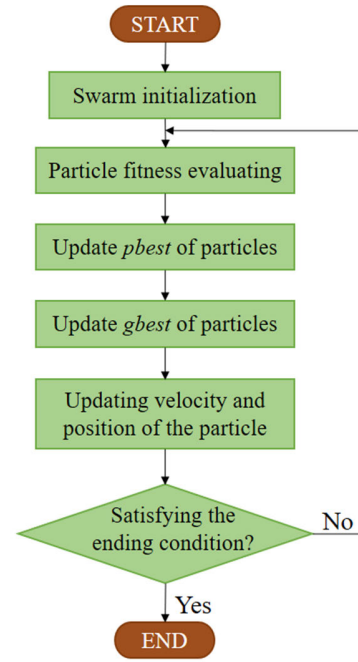
are respectively the velocity and position of the i-th particle after the $(k+1)$-th iteration, and $pbest_i^k$ and $gbest^k$ are respectively, the individual optimal position and global optimal position of the i-th particle (among multiple particles) in the k-th iteration. The optimal solution is output after a specified number of iterations is reached. A flowchart of the process is displayed in Fig. 9.

A. Random Forest Model

The random forest model was proposed by Leo Breiman [28]; it comprises multiple decision tree classifiers and a learning algorithm that involves bagging and random feature sampling. A random forest comprises multiple decision trees; its architecture is displayed in Fig. 10. Each tree makes a prediction, and additional randomness is added. The features are randomly sampled; k samples are taken from the training samples, and k classifiers are trained and input to the original samples. The k samples contain duplicate data, but each tree sample is unique. Finally, voting is performed to determine the prediction result; thus, overfitting is relatively unlikely, and the overall prediction accuracy is high.

B. SVC Model

SVC is a classification method based on SVM [29]. An illustration of the SVC process is presented in Fig. 11. SVC attempts to identify the hyperplane (solid line in Fig. 11) with the greatest distance from the data (margin) as the optimal solution; the dashed lines containing the nearest points are known as the support vectors.

Not all data can be classified using linear classification, separated using a straight line, or placed on a two-dimensional plane. However, the data can be separated on a hyperplane if
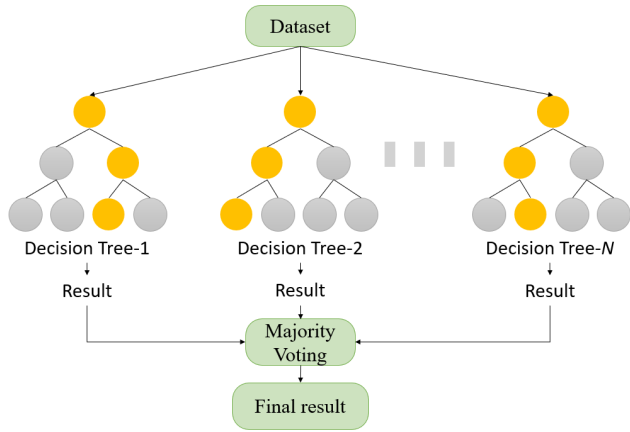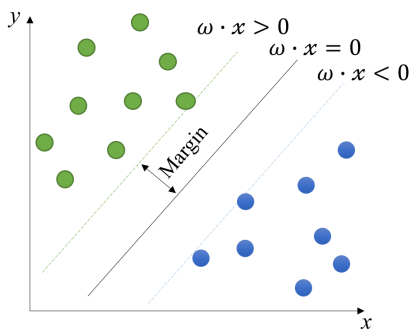
Fig. 10. Random forest architecture.
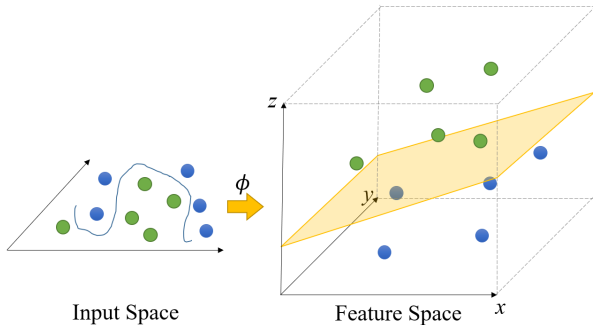

Fig. 11. The concept of SVC.

Fig. 12. Constructing a hyperplane in high-dimensional space.

more spatial dimensions are used. This classification process is nonlinear classification and is illustrated in Fig. 12.

C. AdaBoost Model

AdaBoost is a machine-learning method proposed by Yoav Freund and Robert Schapire [30]. The principle of AdaBoost is that samples misclassified by a previous classifier are used to train a subsequent classifier. AdaBoost is an iterative algorithm in which a new weak classifier is added in each round until a predetermined minimal error rate is reached. Each training sample is assigned a weight indicating the probability it is included in the training set by a certain classifier (Fig. 13). If a sample point has been accurately classified, it has a lower probability of being selected in the next training set, and vice
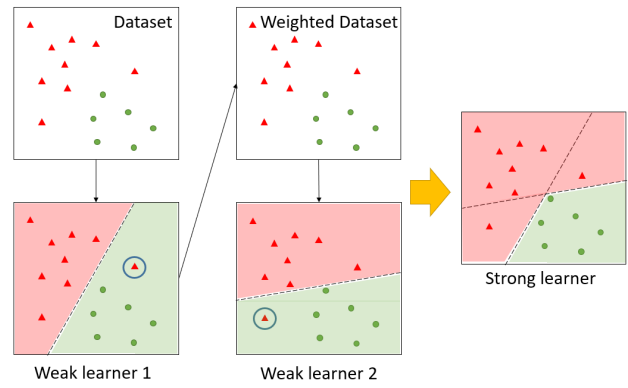

Fig. 13. AdaBoost classification.

versa. Hence, this method focuses on hard-to-classify samples, and overfitting is unlikely.

D. XGBoost Model

XGBoost was proposed by Chen Tianqi [31] and is an extension of gradient boosting that combines bagging and boosting algorithms. XGBoost uses the gradient boosting method illustrated in Fig. 14. Each decision tree is corrected in accordance with previously erroneous predictions to improve model accuracy. Features are randomly sampled to prevent overfitting. The objective function of this model is as follows:

$$\text{Obj} = \sum_{i=1}^{n} \ell(y_i, \widehat{y_l}) + \sum_{k=1}^{K} \Omega(f_k) \tag{16}$$

The objective function comprises both training loss and complexity. $\sum_{i=1}^{n} l(y_i, \hat{y}_i)$ indicates training loss, where l is the loss function, and $\sum_{k=1}^{K} \Omega(f_k)$ indicates complexity. This model is an additive model, and the predicted result after the t-th iteration is as follows:

$$\widehat{y_l}^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \widehat{y_l}^{(t-1)} + f_t(x_i) \tag{17}$$

$\hat{y}_i^{(t)}$ is the prediction result of the i-th sample after the t-th iteration, $\hat{y}_i^{(t-1)}$ is the prediction result of the $(t-1)$-th tree, and $f_t(x_i)$ is the function of the t-th tree. The prediction result of the t-th iteration can be calculated using the prediction result of $t-1$ This is then substituted into the objective function (Obj). $\hat{y}_i$ denotes the training loss of the model. The objective function is further simplified into a quadratic equation to identify the optimal solution. The fundamental principles of this method are presented in [31].

E. MLP Model

MLP, an algorithm based on the human nervous system, is a feedforward artificial neural network that maps a set of input vectors to a set of output vectors. An MLP comprises multiple layers, each with several nodes. These layers can be grouped into an input layer, a hidden layer and an output layer. Specifically, data are received by the input (first) layer, transformed by the hidden layer, and output from the output layer.

Except for the input nodes, each node is a neuron with a nonlinear activation function and a weight that is algorithmically adjusted during training to maximize model accuracy.
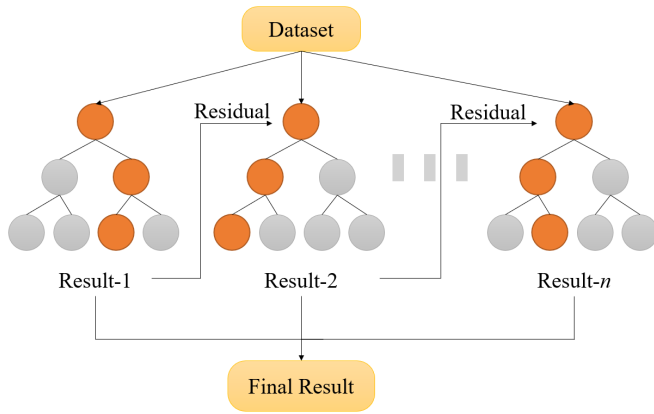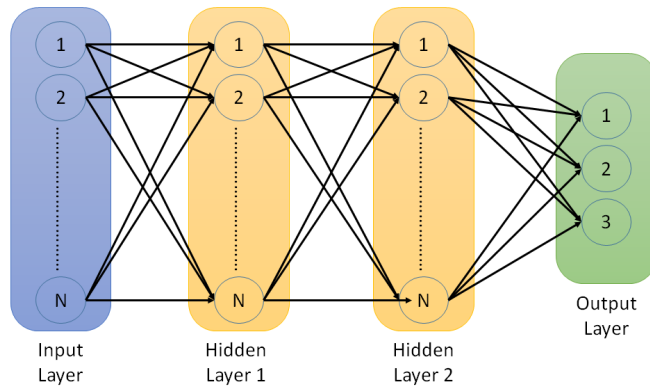
Fig. 14.   XGBoost architecture.

Fig. 15.   MLP architecture.

This approach is well suited to complex problems. The MLP architecture is shown in Fig. 15.

F. Transfer-Learning Model

Transfer learning is a machine-learning method in which the learning results of a base model are transferred to another model such that the new model retains the knowledge of the base model during the learning process. If data are insufficient for training, overfitting may occur. Transfer learning uses the knowledge of the base model to improve the robustness and generalizability of a model, increasing prediction accuracy if the data set is insufficient. Various transfer learning methods have been proposed; that in this study is based on ensemble learning [32], [33]. As shown in Fig. 16, the approach is mainly divided into two stages. In the first stage, the base model is used for training, and in the second stage, the prediction results of the first stage are employed to train the final model. Several models with satisfactory prediction accuracy are first used as the base models, and the prediction results of the base models are selected in accordance with their classification ratios as input to the final model for training. This method ensures that the prediction accuracy of the final model is superior to that of the base models.

The model established in this study (Fig. 17) was produced using transfer learning. First, SMOTE was applied with to augment and balance the training data. The final model was based on MLP, and parameter optimization algorithms were
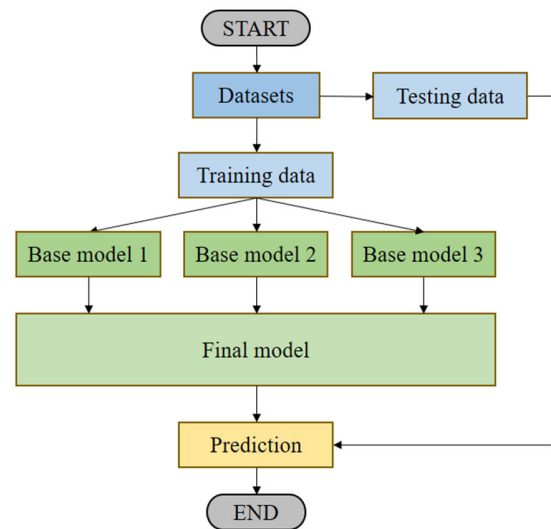

Fig. 16.   Transfer-learning architecture.

applied to optimize the number of model neurons, enabling the model to fit the data and have high accuracy. The data were preprocessed as described in Section II, and k-fold cross-validation was used for model training. The validation data set for final parameter optimization was selected from the SMOTE-augmented training data set. The base models were random forest, SVC, and XGBoost. In the validation data set, the classification results of the base model were represented as probabilities to ensure that the validation and training data sets have the same format. Hence, these data sets can be used as a reference for verifying the model optimization process. The training data output by the base models were input to the final model (the MLP model) for training. The parameters of the final model were optimized by optimization algorithms, and the final model was verified on the validation data. The optimal parameters generated through this process were used for the final model. Finally, the k-fold cross-validation test data were input to the model for prediction. This process was repeated until cross-validation was completed.

IV. THE EXPERIMENTAL RESULTS

In This Section, the Proposed Model Was Compared Against the Modelsintroduced in Section III in Terms of Accuracy; It Was Also Comparedagainst Models From Other Studies That Have Used the Same Data Set as Thepresent Study

A. Comparison of Various Models

The confusion matrix was employed to verify the quality of the classification results for the six models mentioned in Section III and the transfer-learning model integrated with four parameter optimization algorithms.

Fig. 18 presents the confusion matrixes obtained using the five base models (but not the transfer-learning model). The default parameters were used for all models unless otherwise noted. Fig. 18(a) is the confusion matrix for the random forest model. In terms of model parameters, the number of the
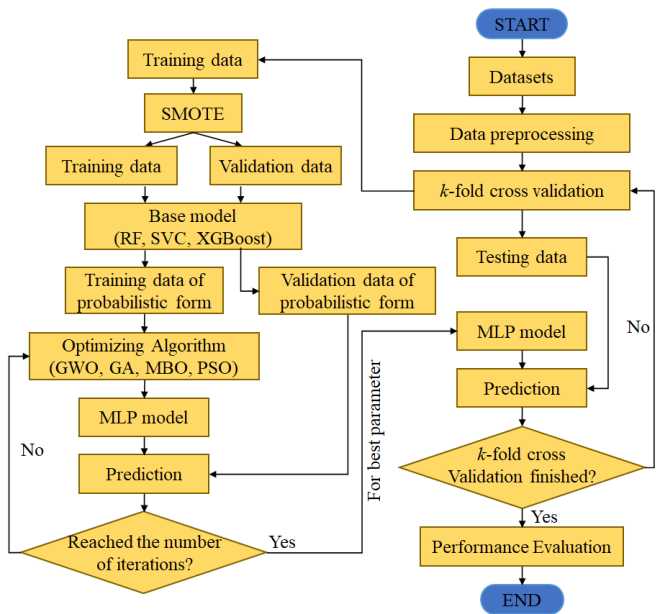
Fig. 17. Flowchart of the transfer-learning model: adding the SMOTE and parameter optimization algorithms.

TABLE II
COMPARISON OF DEMENTIA PREDICTIVE ANALYSES

Model	Accuracy
Random Forest	88.4%
SVC	87.9%
AdaBoost	69.8%
XGBoost	88.7%
MLP	88.1%
Transfer Learning with GWO	90.7%
Transfer Learning with PSO	88.7%
Transfer Learning with GA	89.3%
Transfer Learning with MBO	88.1%

TABLE III
COMPARISON OF DEMENTIA PREDICTIVE ANALYSES WITH OTHERS

Model	Accuracy
Transfer Learning with GWO	90.7%
G.Battineni [17]	68.75%
S.-E. Ryu [6]	85.61%

decision tree was reduced from the default 100 to 10 to reduce training time. The accuracy of the random forest model was 88.4%. Fig. 18(b) is the confusion matrix for the SVC model. The prediction accuracy was 87.9%, which is slightly lower than the default results of random forest.

However, the accuracy for Class 2 was poor. Fig. 18(c) presents the confusion matrix for the AdaBoost model. The accuracy was 69.8%, lowest among the five models. In particular, approximately one-third of the predictions for Class 0 were incorrect. Fig. 18(d) is the confusion matrix for the XGBoost model, which achieved the highest accuracy of 88.7%. However, it had poor accuracy for Class 2. Fig. 18(e) presents the confusion matrix for the MLP model, which had 6 neurons in the input layer and 30, 60, 60, and 20 neurons in the four hidden layers (from layer 2 to layer 5, respectively). The activation function for each hidden layer was ReLU. The output layer had three neurons and the Softmax activation function, which is often used for classification. Although the accuracy of the MLP model was only 88.1%, it had substantially better accuracy for Class 2, which had the least data and for which prediction is the most difficult, than any other model.

The main model established in this study was a transfer-learning model trained on SMOTE-augmented data and optimized with various parameter optimization algorithms. The prediction results of this model are displayed in Fig. 19. The parameter optimization algorithms were used to optimize the model such that the model conformed to various conditions in the data. Compared with the five models in Fig.18, the transfer-learning model yielded much more accurate prediction results except when the MBO algorithm was used for parameter optimization. The prediction result obtained using the GWO algorithm for parameter optimization was highest with accuracy of 90.7%, followed by GA, PSO, and finally MBO.

The accuracy of all models is summarized in Table II, and the findings confirm that the proposed transfer-learning model had superior dementia prediction results than did other models. The GWO algorithm produced the best transfer-learning model with accuracy of 90.7% overall, 95.3% for nondementia (Class 0), 96.9% for dementia (Class 1), and 46% for conversion to dementia (Class 2). The poor results for conversion to dementia class (Class 2) were attributable to the small size of this class; however, this result was substantially better than those of other models. Hence, the model could be effective for predicting dementia.

B. Comparison With Related Studies

Battineni et al. [17] and Ryu et al. [6] applied SVM and XGBoost, respectively, to the data set as that used in this study; hence, their prediction results could be compared with that of the proposed model. Table III reveals that this model had substantially higher overall accuracy than their models; this was attributable to the learning of the base models and the parameter optimization algorithm. The proposed model also had superior accuracy for nondementia, dementia, and conversion to dementia prediction than the models in [6] and [17].

V. DISCUSSION

GA is based on organic evolution and uses the concept of natural selection and survival of the fittest to eliminate genes for optimization. GA can search multiple points to reduce the likelihood of becoming trapped in local optima. In addition, it uses encoding functions for optimization to ensure that the search results are not spatially limited. However, GA cannot guarantee that its final solution is the global optimum. In addition, GA lacks memory; that is, it could search the same points repeatedly, increasing its computational cost. MBO simulates the migration and adaptive behaviors of monarch butterflies to achieve optimization. This algorithm has a simple structure and is easy to implement, and its mathematical model enables each monarch butterfly to fully interact with other butterflies during optimization. However, MBO cannot prevent species from clustering around a local optimum. In addition, the
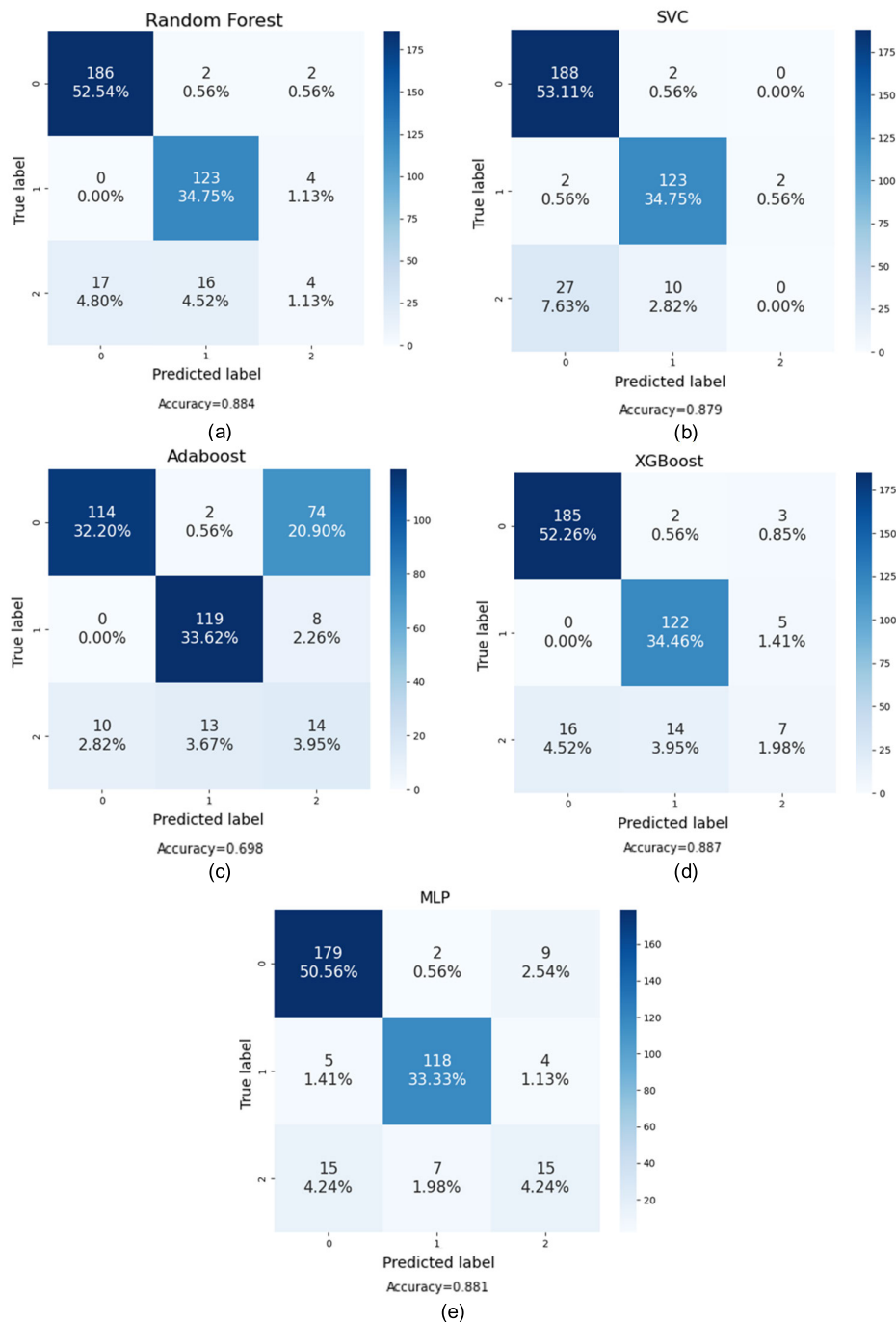
Fig. 18. Confusion matrix for different model. (a) Random Forest, (b) Support Vector Classification, (c) Adaptive Boosting, (d) eXtreme Gradient Boosting, (e) Multilayer Perceptron.

migration involves offspring; hence, regardless of the adaptability of a generated monarch butterfly, its offspring would always inherit this adaptability, thereby decelerating convergence during later calculations. PSO simulates social systems by using multiple particles to search for local optima and then using these local optima to search for the global optimum. PSO has a rapid convergence speed, is simple conceptually, and is easy to implement. In addition, its optimization function can be nondifferentiable or noncontinuous. However, because PSO relies on particles for optimization, it often converges to a local optimum if it finds local extreme values and may

not identify the global optimum. GWO mimics the leadership hierarchy and hunting mechanism of gray wolves in nature, and it identifies better solutions iteratively until a specified number of iterations. It then outputs the optimal solution identified at the final iteration. In each iteration, the problem is divided into multiple subproblems, and GWO searches for the optimal solution to each subproblems. These solutions are ranked to identify the optimal solution for the iteration. Hence, GWO has a faster optimization and better convergence than other optimization algorithms. However, its optimal solution may be a local optimum instead of the global optimum solution
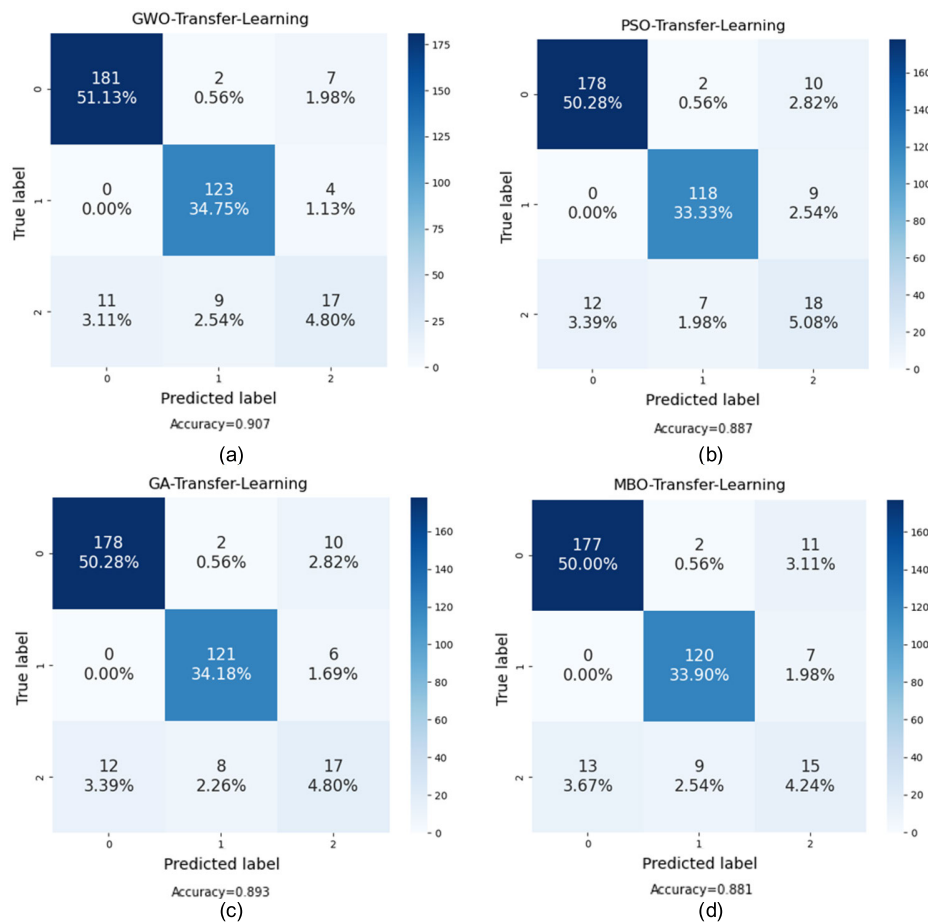
Fig. 19. Confusion matrix for Transfer Learning with different Parameter Optimization Algorithm. (a) GWO, (b) PSO, (c) GA, (d) MBO.

because gray wolves tend to orient toward the location of the leaders of the wolf pack are. Consequently, GWO is worse for global optimization than other algorithms. The results indicate that transfer learning with GWO was superior to that with other optimization algorithms; that is, the data of this study could be best fit with the mathematical model of the GWO optimization process. The experimental results in Section IV also indicate that the GWO yielded better fit. The aforementioned discussion of the advantages and disadvantages of each algorithm reveals that no algorithm can guarantee that the global optimum is identified but all can converge to a local optimal solution. Therefore, GWO generates superior results because it converges to the global optimal solution more effectively than do the other three optimization algorithms, namely GA, MBO, and PSO, which only converge to local optimal solutions and thus produce less accurate results.

Many studies have indicated that earlier dementia diagnosis could improve the effectiveness of pharmaceutical treatment and the cognitive function and ability to perform ADLs of patients after treatment. However, dementia progresses slowly and early symptoms are not obvious; hence, clinical diagnosis is difficult. In general, physicians must eliminate other diseases, such as vitamin B6 deficiency or underactive thyroid, before diagnosing dementia. Consequently, patients have often missed the optimal treatment window by the time that they are diagnosed with dementia. This study provides

a highly accurate model that uses patients and MRI image data to predict whether patients have dementia. The results can serve as a reference for clinical physicians to facilitate dementia identification and diagnosis. In addition, the model can screen patients who may have dementia to enable them to receive treatment as early as possible. The data were labeled as class 0 for nondementia, class 1 for dementia, and class 2 is conversion to dementia. However, patients in class 1 were not classified in accordance with the severity of their dementia, and the ability of the model to predict the various stages of dementia is thus unknown. Conversion to dementia is the optimal time for diagnosing and treating dementia; however, accuracy for classifying class-2 patients was low because class-2 data was insufficient. More data must be collected to produce models that can make more comprehensive predictions.

In summary, the transfer learning with GWO model used by this study produced excellent results, but it also had some limitations. In particular, the model accuracy was limited by the amount of data available. The proposed transfer-learning model of this study was compared with other models in Section IV, namely random forest, SVC, AdaBoost, XGBoost, and MLP. These models were selected for comparisons because they are often applied effectively in various fields. Therefore, these algorithms can be used for a credible comparison.

The system could first be applied to assist experts in identifying patients with dementia (class 1 and 2). Experts could then further inspect patients or use evaluation scales such as CDR or the global deterioration scale (GDS) to diagnose patients. If patients receive a diagnosis of dementia, scales such as the functional assessment staging tool can be used to evaluate the severity of the dementia to formulate a treatment plan. A study [34] listed the challenges of each stage of dementia and the goals of rehabilitation and treatment. A future study could expand the number of classes, aim to predict patients at different stages of dementia, and attempt to predict indices and stages such as GDS or CDR. These stages could include minor and early stages of dementia, MCI before dementia, mild to moderate dementia, and moderate to severe dementia. Rehabilitation for patients with MCI or early can focus on their integration into society. The rehabilitation of patients with moderate dementia who have difficulties with instrumental ADLs could include maintaining their daily activities to avoid further deterioration. In addition, patients with dementia should learn new skills, such as using smartphones, while they can still learn. If patients already have moderate to severe dementia and begin to exhibit abnormal basic ADLs or even suffer from anosognosia, they might lack cognition regarding their own deficits, which would increase the difficulty of rehabilitation. Therefore, rehabilitation for these patients could center on activities that these patients enjoy, including basic life skills, to ensure patients can engage in ADLs. In addition to formulating treatment and rehabilitation plans, the model of this study could be used for regular follow-up to determine disease progress and provide indices for deterioration due to the disease, enabling timely adjustments of treatment and rehabilitation.

VI. Conclusion

Dementia is increasingly prevalent in the context of an aging society. Dementia remains uncurable, and dementia-related neurological degeneration can only be slowed and not stopped. Machine learning could be used to assist health professionals in diagnosing dementia to enable earlier interventions to slow degeneration. This study proposed an effective classification model for dementia prediction by using dementia data from OASIS for predictive analysis. The modified model based on transfer learning was compared with other models. In addition, the model was paired with four parameter optimization algorithms for training, and the results demonstrated that the model had high predictive power and fit the data well. In the future, this model can be used as the primary model for dementia prediction, saving time and serving as a reference in the diagnosis of dementia. Moreover, model instability during training due to SMOTE data augmentation can be mitigated by the use of larger data sets. The proposed system can be used to diagnose dementia and plan occupational therapy regimens.

References

[1] M. Yoneyama, H. Mitoma, N. Sanjo, M. Higuma, H. Terashi, and T. Yokota, "Ambulatory gait behavior in patients with dementia: A comparison with Parkinson's disease," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 8, pp. 817–826, Aug. 2016, doi: 10.1109/TNSRE.2015.2477856.

[2] N. Sharma, M. H. Kolekar, and K. Jha, "Iterative filtering decomposition based early dementia diagnosis using EEG with cognitive tests," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 9, pp. 1890–1898, Sep. 2020, doi: 10.1109/TNSRE.2020.3007860.

[3] P. Durongbhan et al., "A dementia classification framework using frequency and time-frequency features based on EEG signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 5, pp. 826–835, May 2019, doi: 10.1109/TNSRE.2019.2909100.

[4] Z. Xia, T. Zhou, S. Mamoon, and J. Lu, "Recognition of dementia biomarkers with deep finer-DBN," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1926–1935, 2021, doi: 10.1109/TNSRE.2021.3111989.

[5] N. Beebe-Wang, A. Okeson, T. Althoff, and S.-I. Lee, "Efficient and explainable risk assessments for imminent dementia in an aging cohort study," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 7, pp. 2409–2420, Jul. 2021, doi: 10.1109/JBHI.2021.3059563.

[6] S.-E. Ryu, D.-H. Shin, and K. Chung, "Prediction model of dementia risk based on XGBoost using derived variable extraction and hyper parameter optimization," *IEEE Access*, vol. 8, pp. 177708–177720, 2020, doi: 10.1109/ACCESS.2020.3025553.

[7] M. D. Benedetto et al., "Deep networks for behavioral variant frontotemporal dementia identification from multiple acquisition sources," *Comput. Biol. Med.*, vol. 148, Sep. 2022, Art. no. 105937, doi: 10.1016/j.compbiomed.2022.105937.

[8] D. Painuli, S. Bhardwaj, and U. Köse, "Recent advancement in cancer diagnosis using machine learning and deep learning techniques: A comprehensive review," *Comput. Biol. Med.*, vol. 146, Jul. 2022, Art. no. 105580, doi: 10.1016/j.compbiomed.2022.105580.

[9] K. N. Minhad, A. Farayez, K. J. A. Ooi, M. B. I. Reaz, M. A. S. Bhuiyan, and M. H. Miraz, "Sequence prediction algorithm for the diagnosis of early dementia development," in *Proc. Int. Conf. Comput., Electron. Commun. Eng. (iCCECE)*, Aug. 2021, pp. 48–52, doi: 10.1109/iCCECE52344.2021.9534844.

[10] A. H. Alkenani, Y. Li, Y. Xu, and Q. Zhang, "Predicting prodromal dementia using linguistic patterns and deficits," *IEEE Access*, vol. 8, pp. 193856–193873, 2020, doi: 10.1109/ACCESS.2020.3029907.

[11] A. Chaudhary, H. P. Gupta, K. K. Shukla, and T. Dutta, "Sensor signals-based early dementia detection system using travel pattern classification," *IEEE Sensors J.*, vol. 20, no. 23, pp. 14474–14481, Dec. 2020, doi: 10.1109/JSEN.2020.3008063.

[12] Q. Zheng, T. Yu, and F. Wang, "Self-supervised monocular depth estimation based on combining convolution and multilayer perceptron," *Eng. Appl. Artif. Intell.*, vol. 117, Jan. 2023, Art. no. 105587, doi: 10.1016/j.engappai.2022.105587.

[13] J. Ma, Q. Pan, and Y. Guo, "Depth-first random forests with improved grassberger entropy for small object detection," *Eng. Appl. Artif. Intell.*, vol. 114, Sep. 2022, Art. no. 105138, doi: 10.1016/j.engappai.2022.105138.

[14] Q. Hou, L. Liu, L. Zhen, and L. Jing, "A novel projection nonparallel support vector machine for pattern classification," *Eng. Appl. Artif. Intell.*, vol. 75, pp. 64–75, Oct. 2018, doi: 10.1016/j.engappai.2018.08.003.

[15] Y.-L. He, Y. Zhao, X. Hu, X.-N. Yan, Q.-X. Zhu, and Y. Xu, "Fault diagnosis using novel AdaBoost based discriminant locality preserving projection with resamples," *Eng. Appl. Artif. Intell.*, vol. 91, May 2020, Art. no. 103631, doi: 10.1016/j.engappai.2020.103631.

[16] J. Dong, W. Zeng, L. Wu, J. Huang, T. Gaiser, and A. K. Srivastava, "Enhancing short-term forecasting of daily precipitation using numerical weather prediction bias correcting with XGBoost in different regions of China," *Eng. Appl. Artif. Intell.*, vol. 117, Jan. 2023, Art. no. 105579, doi: 10.1016/j.engappai.2022.105579.

[17] G. Battineni, N. Chintalapudi, and F. Amenta, "Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM)," *Informat. Med. Unlocked*, vol. 16, Jan. 2019, Art. no. 100200, doi: 10.1016/j.imu.2019.100200.

[18] D. S. Marcus, A. F. Fotenos, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies: Longitudinal MRI data in nondemented and demented older adults," *J. Cognit. Neurosci.*, vol. 22, no. 12, pp. 2677–2684, Dec. 2010, doi: 10.1162/jocn.2009.21407.

[19] M. Roheger, H. Xu, M. T. Hoang, M. Eriksdotter, and S. Garcia-Ptacek, "Conversion between the mini-mental state examination and the Montreal cognitive assessment for patients with different forms of dementia," *J. Amer. Med. Directors Assoc.*, vol. 23, no. 12, pp. 1986–1989, Dec. 2022, doi: 10.1016/j.jamda.2022.03.018.

[20] Y.-C. Wang and C.-H. Cheng, "A multiple combined method for rebalancing medical data with class imbalances," *Comput. Biol. Med.*, vol. 134, Jul. 2021, Art. no. 104527, doi: 10.1016/j.compbiomed. 2021.104527.

[21] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Adv. Eng. Softw.*, vol. 69, pp. 46–61, Mar. 2014, doi: 10.1016/j. advengsoft.2013.12.007.

[22] H. Rezaei, O. Bozorg-Haddad, and X. Chu, "Grey wolf optimization (GWO) algorithm," in *Advanced Optimization by Nature-Inspired Algorithms*. Singapore: Springer, 2018, pp. 81–91, doi: 10.1007/978-981-10-5221-7_9.

[23] C. Lin, "An adaptive genetic algorithm based on population diversity strategy," in *Proc. 3rd Int. Conf. Genetic Evol. Comput.*, Oct. 2009, pp. 93–96, doi: 10.1109/WGEC.2009.67.

[24] A. Lambora, K. Gupta, and K. Chopra, "Genetic algorithm—A literature review," in *Proc. Int. Conf. Mach. Learn., Big Data, Cloud Parallel Comput. (COMITCon)*, Feb. 2019, pp. 380–384, doi: 10.1109/COMITCon.2019.8862255.

[25] G.-G. Wang, S. Deb, and Z. Cui, "Monarch butterfly optimization," *Neural Comput. Appl.*, vol. 31, no. 7, pp. 1995–2014, Jul. 2019, doi: 10.1007/s00521-015-1923-y.

[26] X. Cui, Z. Chen, and F. Yin, "Differential evolution and local search based monarch butterfly optimization algorithm with applications," *Int. J. Comput. Intell. Syst.*, vol. 12, no. 1, p. 149, 2018, doi: 10.2991/ijcis.2018.25905188.

[27] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. ICNN Int. Conf. Neural Netw.*, Nov. 1995, pp. 1942–1948, doi: 10.1109/ICNN.1995.488968.

[28] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[29] J. Kang, X. Han, J. Song, Z. Niu, and X. Li, "The identification of children with autism spectrum disorder by SVM approach on EEG and eye-tracking data," *Comput. Biol. Med.*, vol. 120, May 2020, Art. no. 103722, doi: 10.1016/j.compbiomed.2020.103722.

[30] Y. Freund and R. E. Schapire, "A short introduction to boosting," in *Proc. 6th Int. Joint Conf. Artif. Intell.*, 1999, pp. 1401–1406.

[31] T. Chen and C. Guestrin, "XGBoost," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[32] P. K. Roy and A. Kumar, "Early prediction of COVID-19 using ensemble of transfer learning," *Comput. Electr. Eng.*, vol. 101, Jul. 2022, Art. no. 108018, doi: 10.1016/j.compeleceng.2022.108018.

[33] R. C. Joshi, R. Mishra, P. Gandhi, V. K. Pathak, R. Burget, and M. K. Dutta, "Ensemble based machine learning approach for prediction of glioma and multi-grade classification," *Comput. Biol. Med.*, vol. 137, Oct. 2021, Art. no. 104829, doi: 10.1016/j.compbiomed.2021. 104829.

[34] Y. Maki, T. Sakurai, J. Okochi, H. Yamaguchi, and K. Toba, "Rehabilitation to live better with dementia," *Geriatrics Gerontol. Int.*, vol. 18, no. 11, pp. 1529–1536, Nov. 2018, doi: 10.1111/ggi.13517.