IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 January 2024, accepted 13 February 2024, date of publication 21 February 2024, date of current version 7 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3368067

RESEARCH ARTICLE

ERDeR: The Combination of Statistical Shrinkage Methods and Ensemble Approaches to Improve the Performance of Deep Regression

ZARI FARHADI[1,2], MOHAMMAD-REZA FEIZI-DERAKHSHI[2], HOSSEIN BEVRANI[1], WONJOON KIM[3], (Member, IEEE), AND MUHAMMAD FAZAL IJAZ[4]

[1]Department of Statistics, Faculty of Mathematics, Statistics and Computer Sciences, University of Tabriz, Tabriz 5166616471, Iran
[2]Computerized Intelligence Systems Laboratory, Department of Computer Engineering, Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz 5166616471, Iran
[3]Division of Future Convergence (HCI Science Major), Dongduk Women's University, Seongbuk-gu, Seoul 02748, South Korea
[4]School of IT and Engineering, Melbourne Institute of Technology, Melbourne, VIC 3000, Australia

Corresponding authors: Mohammad-Reza Feizi-Derakhshi (mfeizi@tabrizu.ac.ir) and Wonjoon Kim (wjkim@dongduk.ac.kr)

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT: Ministry of Science and ICT) (No. RS-2023-00239877).

ABSTRACT Ensembling is a powerful technique to obtain the most accurate results. In some cases, the large number of learners in ensemble learning mostly increases both computational load during the test phase and error rate. To solve this problem, in this paper we propose an Ensemble of Reduced Deep Regression (ERDeR) model, which is a combination of Deep Regressions (DRs), shrinkage methods, and ensemble approaches. The framework of the proposed model contains three phases. The first phase includes base regressions in which parallel DRs are used as learners. The role of these DRs is to extract features of input data and make prediction. In the second phase, to automatically reduce and select the most suitable DRs, shrinkage methods such as Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net (EN) are employed. These models are compared with the non-shrinkage model. The last phase is ensemble phase, which consists of three different ensemble methods namely Multi-Layer Perceptron (MLP), Weighted Average (WA), and Simple Average (SA). These ensemble methods are used to aggregate the remaining learners from previous steps. Finally, the proposed model is applied to Monte Carlo simulation data and three real datasets including Boston House Price, Real Estate Valuation and Gold Price per Ounce. The results show that after applying the shrinkage methods the error rate is significantly reduced and the model accuracy is increased. Accordingly, the results of combining shrinkage methods and ensemble approaches not only decreased the computational load during test phase, but also increased the model accuracy.

INDEX TERMS Deep learning, convolutional neural network, shrinkage methods, ensemble learning, LASSO, elastic net.

I. INTRODUCTION

Machine learning-based techniques have been applied in several situations, including economic areas [1] object detection, forecasting stock prices, and medical profession [2]. They can be divided into single, ensemble, and hybrid learning techniques. Deep learning (DL) is a particular branch of ML that has achieved outstanding achievements in various fields such as pattern recognition, image processing, speech generation, signal processing, and forecasting. A comparison between traditional machine learning and deep learning algorithms shows that deep learning algorithms have many advantages. One of these advantages is the use of different layers and neural networks to make a deep network with accurate analysis.

Convolutional Neural Network (CNN) is a special type of Artificial Neural Network (ANN) that makes a powerful deep learning network by increasing the number of layers and

The associate editor coordinating the review of this manuscript and approving it for publication was Juan A. Lara.

nodes. CNN models learn to automatically extract features using filters and convolutional layers. These models have different structures, including 1D, 2D, and 3D-CNN, which can be used for time series, image processing, and video recognition, respectively. In addition, 2D-CNN with a large number of convolutions and hidden layers as well as millions of parameters can learn complex patterns; however, it may not be suitable for 1D signals. Therefore, the 1D-CNN has recently been developed as an alternative method and a modified version of 2D-CNN, which has many applications due to its low computational complication and relatively shallow architecture [3].

Overfitting is a tremendous challenge for researchers in the field of ML and DL. If the complexity of deep learning and machine learning models is not properly designed, they can be vulnerable to overfitting. In ML algorithms such as Random Forest (RF), increasing the number of trees causes overfitting, which creates problems in training phase as well as testing phase. This issue was investigated by Farhadi et al. [4], [5]. In addition, in DL algorithms such as CNN, which have a layered structure, the depth of network increases by adding layer. Although excessive depth may increase the accuracy, it can increase complexity and computational load as well as overfitting in the model. To avoid overfitting, there are different statistical approaches such as regularization and shrinkage methods; these approaches can reduce overfitting and at the same time improve model performance and accuracy. In the present paper, we use shrinkage methods to reduce the number of learners. For example, LASSO and EN can be utilized as regularization methods to reduce the number of nodes and relationships between extracted features in fully connected layers and convolutional layers in CNN, respectively. This problem was discussed in paper [6]. Using regularization methods can reduce overfitting and enhance the model's accuracy as well. Accordingly, the main purpose of the present paper is to limit the complexity of network by applying shrinkage methods to reduce the computational load and avoid overfitting.

Nowadays, reducing computational load has also become a major challenge for improving ML algorithms because the large number of learners in ensemble-based algorithms can cause overfitting. Reducing the number of learners can be beneficial to control overfitting. Some tools such as shrinkage methods can be used to reduce overfitting in ML, DL algorithms and hybrid methods [7]. In order to overcome the overfitting and computational load as well as to improve the prediction accuracy, the concept of ensemble learning has been recently introduced by researchers to solve both classification and regression problems [8], [9].

An ensemble prediction model consists of a set of individually trained base models, e.g., neural networks and decision trees, whose outputs are combined to make a prediction [8], [10], [11], [12]. By taking advantage of model complementarity, ensemble prediction model can provide more stable and accurate predictions than the conventional single prediction model [13]. Because of its appealing prediction performance, ensemble prediction approach has received greater attention and become a sought-after research topic in many fields, particularly in data classification [14], disease diagnosis such as diabetes [15], [16], [17], [18], [19], pattern recognition [20], [21], [22], image processing [23], [24], [25], [26], [27], speech generation [28], [29], [30], [31], and signal processing [32], [33].

In recent years, beyond image and signal processing, scientists have applied deep learning algorithms in other fields such as housing price prediction [34], disease recognition [35], [36], fault diagnosis [37], and prediction of traffic flow [38], [39]. Deep learning algorithms [40] such as CNN, Recurrent Neural Network (RNN), and Long Short Term Memory (LSTM) or a hybrid of these algorithms, e.g. CNN-RNN [41], [42], and CNN-LSTM [39], [43], [44], [45] are widely used. In order to predict the traffic speed in a specific area of Hong Kong, Cao et al. [39] modeled applications of deep learning based on a hybrid of CNN and Long-Short-term Memory (LSTM) called CNN-LSTM. In addition, that hybrid model was compared with SVR, LASSO, RF, MLP, and LSTM. As a result, in this proposed hybrid model outperformed the other algorithms. Canizo et al. [41] used DL-based methods to detect supervised multi-time series anomaly in multi-sensor systems, in which the CNN and RNN were combined in different ways and created a hybrid model called CNN-RNN.

The combination of several learners' outputs makes ensemble models that generate a single output. These models are designed in such a way that their basic learners have the same structure [46], [47], [48], [49]. Furthermore, another important issue addressed in the ensemble models is how to integrate the models to determine final decision. These models are built using different types of ML algorithms that can perform differently in various conditions. Some of these algorithms are RF and Bagging, in which decision trees are their basic learners. The other algorithm is the ensemble of Support Vector Machine (SVM) [50], ensemble of Neural Network [51], and ensemble of k-Nearest Neighbor (KNN) [52] from which SVM, ANN and KNN are employed as a learner, respectively. In [53], a multi-level feature selection algorithm based on LASSO coefficient threshold (Coe-Thr-Lasso) was proposed. In the proposed algorithm, Lasso-based feature selection was used to remove features with redundancy and weak correlation. After reducing the features with weak correlation, three machine learning algorithms, including Logistic Regression (LR), RF, and SVM were combined with the proposed algorithm. The results showed that the proposed Coe-Thr-Lasso algorithm outperformed three ML algorithms namely LR, RF, and SVM. In [54], a new enhanced feature selection method was proposed to improve the training model and classification performance of the model. This improvement was achieved based on the concept of regularization in which the selection of the best features was considered before training the model under any propagation environment. The adoption of regularization leaded to a high Total Explained Variance (TEV) during the process of kernel Principal Component Analysis (k-PCA). The selected features were the input of ML algorithms such as KNN and SVM. In other words, the enhanced feature selection method was combined with ML algorithms. The proposed model reduced the dimension of features and, generally, enhanced the ML classification performance [55]. In [56], a novel ensemble of 3D-DenseNet was proposed to boost the performance of dementia detection model. It was constructed by varying hyper-parameters and architecture around the optimal values for base 3D-DenseNets. In this model, first, dense connections were introduced to maximize the information flow, where each layer connects with all subsequent layers directly. Then probability-based fusion method was employed to combine 3D-DenseNets with different architectures. One of the most advantages of ensemble methods employed in this proposed model is the reduction of the misclassification risk of a single classifier.

In [57], a method called the Random Ensemble Deep Spatial (REDS) approach was proposed to predict spatial data. This procedure used random Fourier features as inputs to an extreme learning machine (a deep neural model with random weights), and with calibrated ensembles of outputs from this model based on different random weights, it provided a simple uncertainty quantification. The REDS method was demonstrated on simulated data and on a classic large satellite data set. The purpose of the study [58] is to combine multiple single machine learning models with integrated learning algorithms and propose an SMC retrieval method based on multiple differentiated models under a stacking integrated learning architecture. First, 19 factors, including: radar backscattering coefficient, vegetation index, and drought index, that affect SMC were extracted from SENTINEL-1, LANDSAT, and terrain factors. Those with the highest importance scores were selected as retrieval factors using the Boruta algorithm combined with four single machine learning methods—classified regression tree, random forest, gradient boosting decision tree (GBDT), and extreme random tree. In addition, the two stacking ensemble models using least absolute shrinkage and selection operator (LASSO) and the generalized boosted regression model (GBM) were tested and applied to build the most reliable and accurate estimation model. In [59], two single-based and stacked ensemble-based machine learning models were employed to speed up the parameter estimations of wireless sensor network with highly accurate outcomes. Adaboost was superior over other models (Elastic Net, SVR) in single-based models. Stacked ensemble models achieved best results for the WSN parameter prediction compared to single-based models.

Recent advances in deep learning show that using ensembles of CNNs can improve the performance of prediction [60], [61], [62]. Although the ensemble of CNNs can have better performance than single models, this research path remains significantly unknown in the literature [63]. The main reason is that the ensemble of CNNs needs more time for training and diverse combination for prediction. For example, in the post-selection boosting random forest (PBRF) [64] algorithm, decision tree is used as a learner, while in the present study, we use Deep CNN. Additionally, after applying the LASSO method on the decision trees in PBRF, Simple Average (SA) was used to aggregate the trees. In [5], to improve the accuracy of RF, the combination of clustering algorithm (K-means) and shrinkage methods (LASSO, Elastic Net, and Group Lasso) were used. This proposed framework, called ECAPRAF, is a machine learning ensemble model that automatically reduces the number of trees in RF by shrinkage methods; while in our proposed model several different ensemble methods with various structures were used. Table 1 summarizes the various ML and DL approaches for ensemble learning approaches [65].

In this article, a hybrid model called Ensemble of Reduced Deep Regression (ERDeR) is presented in which a combination of DRs, ensemble approaches, and shrinkage methods are used to increase accuracy and improve the performance of the model. The DRs, which are composed of Deep CNN architecture, are operated as base learners that have the same structure.

Additionally, in our proposed model, three different ensemble approaches and two shrinkage methods are used to ensemble and reduce DRs, respectively. They are compared with the non-shrinkage model. The architecture consists of three main phases, along with the fusion phase. The first phase includes base regressions which is made of several parallel DRs as a learner. The number of DRs is set to 30, 40, and 50.

These are responsible for the feature extraction and prediction of the input dataset. Each of the DRs is formed of convolution, pooling, and fully-connected layers, which are used to extract features, reduce the dimension of input data, and predict, respectively. The next phase is fusion, which involves concatenating all trained DRs in the base regression phase and ensembling the remaining DRs of the shrinkage phase. Since no special operations are performed in the fusion phase, it is not considered as the main phase.

In this model, among the constructive phases, two phases are especially important. The first phase is the shrinkage phase. In this phase, LASSO and Elastic Net (EN) methods are used to reduce the number of DRs in the model. They are compared with the non-shrinkage model in which removing the DRs are not accomplished.

In addition, these methods improve the model by selecting appropriate DRs. In LASSO method, the sum of absolute values of regression coefficients is used as a penalty function. In the case of EN, the linear combination of the sum of squared regression coefficients and the sum of absolute values of the regression coefficients are used. The EN has a much better performance than LASSO.

It should be noted that DRs are automatically removed without any initial selection. This is associated with reducing

TABLE 1. The related work of the machine and deep learning approaches for ensemble learning and Deep regressions.

Ref.	Approach	Objective	Challenges of the Approach
[66] , [67]	Ensemble of CNN	An ensemble-based classification model was developed using three Convolutional Neural Network (CNN) architectures, namely Inception v3, Xception and DenseNet-169 pre-trained on ImageNet dataset for Pap stained single cell and whole-slide image classification.	An automated screening framework was proposed for the detection of cervical cancer. Although simple fusion schemes like majority voting, weighted averaging, etc., have been used in literature, they do not consider the confidence in the predictions of a classifier while computing the predictions.
[68]	Deep Regression with Shrinkage Loss	To estimate target positions, regression trackers directly learn a mapping from regularly dense samples of target objects to soft labels, which are usually generated by a Gaussian function.	To balance training data, a novel shrinkage loss was proposed to penalize the importance of easy training data. Additionally, the residual connections were applied to fuse multiple convolutional layers as well as their output response maps.
[69]	Robust Optimization for Deep Regression	Reducing the influence of outliers in the model fitting process by proposing a loss function that is robust to outliers.	A regression model with ConvNets was proposed that achieved robustness to such outliers by minimizing Tukey's biweight function, an M-estimator robust to outliers, as the loss function for the ConvNet.
[70]	Vanilla Deep Regression	Comprehensive analysis of deep regression techniques	A systematic evaluation and statistical analysis of vanilla deep regression, i.e. convolutional neural networks with a linear regression top layer.
[71]	XGBoost and deep learning	Establishing a variety of machine learning techniques, such as white-box machine learning, deep learning, and ensemble learning to determine the solubility of H_2S in ionic liquids (ILs).	Employing deep belief network (DBN), and extreme gradient boosting (XGBoost) in order to construct genetic programming (GP), group method of data handling (GMDH).
[72]	CNN, VGG16, and VGG19	Applying a variety of image processing techniques including the data augmentation technique to increase the number of data and solve the overfitting problem of the model, and proposing ensemble learning model of custom CNN, Transfer Learning, and CNN-Machine Learning (ML) classifier techniques	applying a max voting ensemble technique in combination with adaptive weighted average ensemble models to reduce the dispersion of predictions

errors, increasing accuracy and improving model performance. Another important phase is an ensemble which consists of various ensemble approaches used to aggregate the previous steps. The ensemble approaches include MLP, WA, and SA. Each of the mentioned ensemble methods performs aggregation by applying shrinkage methods and without them. We call the non-shrinkage models MLP-NON, WA-NON, and SA-NON. One of these methods is Multi-Layer Perceptron which is made of several fully-connected layers and activation functions. This method applies a non-linear transformation on output. Another method is Weighted Average which applies a linear transformation on output. This method is made up of the combination of single fully-connected layer and activation functions. In principle, this combination creates a type of weighted averaging that has better performance than other ensemble methods, such as MLP and WA. Another method, which is the simplest method compared to the other two methods, is the SA. That only performs SA. Finally, the results of all three models are compared with shrinkage models and non-shrinkage model. The number of different DRs is considered as well.

The main objective of our proposed model is to design a combination of DRs, shrinkage methods, and different ensemble approaches to increase the model's accuracy and reduce computational load in the test phase and overfitting as well as remove ineffective learners by shrinkage methods. So, the main strength of the present study is the use of a combination of DRs, ensemble approaches, and shrinkage methods for learning, aggregating, and reducing learners, respectively. The ERDeR model aggregates the DRs by MLP, WA, and SA ensemble approaches after applying shrinkage methods. They are compared with non-shrinkage models in which no DRs are removed. Moreover, to improve the proposed model, the shrinkage methods such as LASSO and EN are applied to reduce the number of DRs. After shrinking, the remaining DRs are aggregated using ensemble methods described earlier.

The main contributions of our study are as follows:

- Using DRs with 1D-CNN structure as a learner to show that deep learners can be used instead of decision trees or other algorithms to improve the performance of model.
- Using shrinkage methods to reduce the number of learners and computational load in the test phase as well as improve the performance of the ERDeR model.
- Using different structures of ensemble learning model to aggregate the remaining learners with and without applying shrinkage methods.
- Our method performs well against state-of-the-art structure of deep regressions and ensemble models. We succeed to reduce the number of deep regressions by shrinkage methods and improve the model's performance.

The rest of the paper is organized as follows: Section II presents an explanation of the related methods briefly. Section III provides a flowchart and summary of the proposed method. In section IV, the performance of the proposed model is evaluated and analyzed by a simulation study and three real datasets. Finally, Section V includes the conclusion and future works.

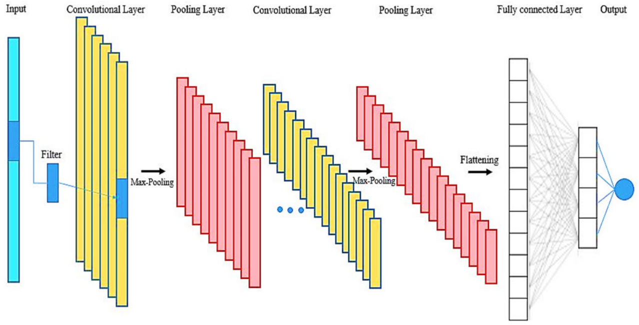


FIGURE 1. **1D Convolutional neural network structure.**

II. METHODOLOGY

A. ONE-DIMENSIONAL CONVOLUTIONAL NEURAL NETWORK

CNN is the most well-known method in deep learning. It is a feed-forward and multi-stage neural network with two types of layers consisting of Convolution and fully-connected layers similar to MLP neural network. The 1D convolutions and sub-sampling both occur in convolutional layers and fully-connected part which perform the classification and prediction operation. In this study, Deep Convolution Neural Network is used for regression, which is shown in Figure 1. In the convolution and pooling layers, different kernel sizes are used to extract the features of the input dataset. The feature map of the previous layer is the input for the next convolutional layer along with the kernel. Finally, the nonlinear activation function forms the output feature map.

The CNNs are mainly trained in a supervised way by the back propagation (BP) algorithm [73]. It can be said that the network parameters are updated using this algorithm. The update is done by calculating the weighted gradient of the whole loss function layers. This update will be continued until a certain stopping criterion is obtained. In the BP algorithm, several gradient descent optimization methods can be used including Stochastic gradient descent [74], AdaGrad [75], RMsprop, and Adam [76]. In this paper, the Adam gradient descent is used. The purpose is to minimize the loss function by reducing the share of network parameters [77].

1) CONVOLUTION LAYER

The 1D-CNNs perform convolutional operations on input to obtain one-dimensional features in which various filters extract different features. In other words, the convolutional layers convolve the input and learn to extract features based on the convolution operator that will be used by the fully-connected layer for classification or regression. Each convolutional layer includes filters that identify the features and extract them from input data. One filter corresponds to one feature map in the next layer. The output of the convolutional layers is the input of the next layer [78]. The calculation process in the one-dimensional convolutional layer is as follows:

$$X_j^l = f\left(\sum\nolimits_{i\epsilon M_j} X_i^{l-1} * w_{ij}^l + b_j^l\right) \tag{1}$$

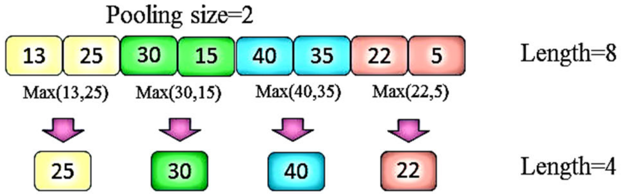


FIGURE 2. **1D Max pooling.**

where M_j refers to a set of input feature maps, l represents the l-th layer, X_i^{l-1} indicates the i-th input of the feature map of $(l-1)$-th layer, w_{ij}^l shows the number of kernel convolutions that is ij-th feature map, X_j^l indicates the j-th feature map of l-th layer, f (.) is a non-linear activation function in which the $*$ operator is used to perform the convolutional operations, b is the kernel bias corresponding to the kernel.

2) POOLING LAYER

After the convolutional operation in the convolution layers, the dimension of feature maps is increased. Therefore, to reduce the dimension of the feature map as well as control overfitting, pooling layer is intended as a reducer to decrease the parameters of the whole network. Although several pooling methods such as max-pooling (calculating the maximum value for each patch of the feature map) and average-pooling (calculating the average value for each patch on the feature map) can be used in CNN, the most commonly used pooling layer is max-pooling which performs the maximum operation over the input features. Figure 2 indicates how to calculate the 1D max-pooling layer. In this study, max-pooling is used to decrease the computational load. Pooling layers are calculated as follows:

$$X_j^l = f\left(\beta_j^l .down\left(X_j^{l-1}\right) + b_j^l\right) \tag{2}$$

where β_j^l and b_j^l represent the multiplicative bias and additive bias of the l^{th} layer, respectively. Also, $down$ (.) represents the subsampling or pooling function.

3) FULLY CONNECTED LAYER

The last part of CNN is called fully-connected (FC) layers, which is a particular and main part of CNN. For doing classification or regression, the output of the previous layers which makes the input of fully-connected part is converted into 1D feature vectors. This part is constructed from various hidden layers, neurons, and activation functions. Additionally, the loss function can include mean-squared error (MSE), cross entropy, and logarithmic [79]. In this study, MSE of the cost function is used.

B. ENSEMBLE LEARNING

Ensemble strategy is a term used to describe the methods in which multiple base learners are combined to make a joint decision and provide higher accuracy and better results than a single learner [80]. These learners can be any type of classification algorithm. For example, in the RF which is one of the most famous tree-based XGboost ensemble learning algorithms, the decision tree is used as a learner. In this study, CNN is applied as a base learner. For the final prediction, there are several ensemble learning approaches such as majority voting, average voting, and stacking for classification tasks [81]. In addition, the simple averaging [82] and weighted averaging [83] strategies are applied for regression tasks. In this research, we use MLP, SA, and WA. SA is one of the simplest and most effective implementation methods used in neural networks. It is calculated in a way that posterior values are produced by averaging the predicted values of all base learners. This is accomplished by compensating for the error of a single learner by other learners to lead to stronger classification performance. Another method is weighted averaging which can be obtained by minimizing a loss function. It means that the classifiers which have better performance are assigned larger weights by minimizing the loss function. The final ensemble prediction is obtained by averaging from the optimal weights of classifiers [84]. In other words, it is possible to calculate the optimal weights of the network by using several fully-connected layers and then ensemble them to achieve an optimal ensemble classifier [51].

C. LASSO

One of the most popular shrinkage methods, introduced by Tibshirani [85], is LASSO regression. This method creates a suitable model by selecting effective variables and removing ineffective ones. In this research, this method is used to reduce the number of parallel DRs (learners) by applying a constraint on the absolute value of regression coefficients [86]. This causes the regression coefficients to tend toward zero or become exactly zero. The applied constraint depends on a tuning parameter λ whose value is obtained based on cross-validation. As the value of parameters increases, a large number of coefficients become zero. In some cases, this removes the effective variables from the model leading to overfitting and error increase. Choosing an appropriate λ can improve the model performance.

Suppose $(\mathbf{X}, Y)$ is a dataset so that $\mathbf{X} = \left(x_1, \ldots, x_p\right)'$ is the prediction variable and Y is the response variable. To estimate LASSO, the ℓ_1-norm is used based on the penalized least squares, which is defined as follows:

$$\hat{\beta}^{lasso} = \underset{\beta\epsilon\mathbb{R}^p}{argmin}\left\{\frac{1}{2N}\|Y - X\beta\|_2^2\right\}\ s.t\ \|\beta\|_1 \leq t \tag{3}$$

where t and β are the tuning parameter and regression coefficient, respectively. This equation can be converted to the penalized least square function whose Lagrangian form will be as follows:

$$L(\beta, \lambda) = \underset{\beta\epsilon\mathbb{R}^p}{argmin}\left\{\frac{1}{2N}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1\right\} \tag{4}$$

where $\|\beta\|_1 = \sum_{j=1}^{p}\left|\beta_j\right|$ and $\left\|Y - \beta^T X\right\|_2^2 = \sum_{i=1}^{n}\left(Y - (X\beta)_i\right)^2$ are ℓ_1-norm under β and loss function, respectively. λ is the tuning parameter, which controls the extent to which the coefficients are penalized towards zero ($\lambda \geq 0$). When λ is zero, the values of estimated coefficients are identical to the standard regression coefficients. In addition, λ is obtained based on 10-fold cross validation.

D. ELASTIC NET

Although the LASSO can simultaneously shrink and select the features, it has some limitations and is not always suitable [87]. Therefore, Zou and Hastie [88] developed the EN EN in 2005, which is an excellent regression method combining the advantages of LASSO and Ridge. This method can improve the constraints as a two-step method, so that in the first step, it performs Ridge-type contraction (ℓ_2-norm) and then LASSO-type thresholding (ℓ_1-norm). In principle, it is a convex linear combination of Ridge and LASSO methods in a way that if $\alpha = 1$ and $\alpha = 0$, it is converted to Ridge and LASSO regression, respectively. As a result, this method has both the Ridge and LASSO regression properties.

Suppose $(\boldsymbol{X}, Y)$ is a dataset so that $\boldsymbol{X} = \left(x_1, \ldots, x_p\right)'$ is the prediction variable and Y is the response variable. To estimate the EN, the combination of ℓ_2 and ℓ_1-norm are used based on the penalized least squares to solve the optimization problem, which is defined as follows:

$$\hat{\beta}^{elastic} = \underset{\beta\epsilon\mathbb{R}^p}{argmin}\left\{\|Y - X\beta\|^2\right\}\ s.t\ (1 - \alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1 \leq t \tag{5}$$

where $(1 - \alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1$ is called EN penalty so that α is equal to $\frac{\lambda_2}{\lambda_1 + \lambda_2}$. For all constant and non-negative values of λ_1 and λ_2, the Lagrangian form is defined as follows:

$$L(\lambda_1, \lambda_2, \beta) = \underset{\beta\epsilon\mathbb{R}^p}{argmin}\left\{\frac{1}{N}\|Y - X\beta\|^2 + \lambda_2\|\beta\|_2^2 + \lambda_1\|\beta\|_1\right\} \tag{6}$$

where $\|\beta\|_1 = \sum_{j=1}^{p}\left|\beta_j\right|$ and $\|\beta\|_2^2 = \sum_{j=1}^{p}\beta_j^2$ show the ℓ_1 and ℓ_2-norm, respectively. Also, λ_1 and λ_2 are tuning parameters.

III. STRUCTURE OF PROPOSED MODEL

A. THE CONSTRUCTION OF ENSEMBLE OF CONVOLUTIONAL NEURAL NETWORKS

The purpose of ERDeR model, which is made of the combination of deep regression, ensemble approaches and shrinkage methods, is to predict and improve the performance of the designed model. This ensemble model consists of three phases including base regression, shrinkage, and ensemble. All DRs receive a dataset as an input and give the corresponding prediction. All DRs have the same structure. They are made of deep CNN which extract features and predict. Each network has several convolutional layers made up of different numbers of filters. The tanh activation function is applied to the results of convolutional operation. Each convolutional network has several pooling layers with a fixed pooling

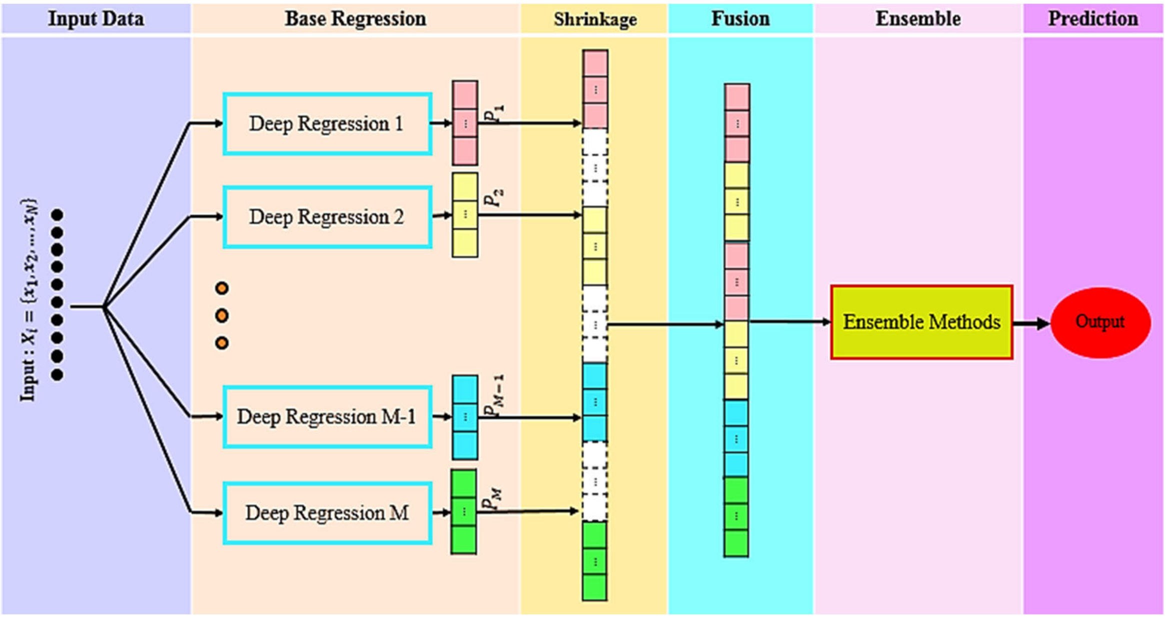


FIGURE 3. The structure of proposed ERDeR model.

size. After convolving and pooling in each DR, two fully-connected layers with various numbers of units are used, each of which is responsible for prediction. By applying the ReLU activation function on the hidden layer and the Linear activation function on the output layer, the prediction value of each network is calculated.

B. THE AGGREGATION STRATEGY

The next step is a strategy to ensemble the results of DRs and reduce their number, which includes two phases. Three options are used to ensemble including Multi-Layer Perceptron, WA and SA. They will be explained in the next sections in details. Another option used to reduce the number of DRs is shrinkage methods such as LASSO and EN. These methods in combination with each other produce new models that are compared with the non-shrinkage model where the shrinkage method is not applied on DRs. In addition to the stated phases, the proposed model also includes a fusion phase that performs concatenation on DRs. Due to the low importance of this phase, we leave out its details. The whole structure of the ERDeR model is shown in Figure 3. The details of the inner layers will be explained in the next subsections.

C. MODEL 1: MLP ENSEMBLE

The MLP ensemble model makes up of parallel DRs. These base regressions are made of convolutional layers with different hyperparameters. The details of the parameter setting are given in Table (2).

As can be seen in Figure 4, in the fusion phase, concatenation is used to merge DRs. In other words, in fusion phase, the extracted deep features of input data are concatenated. After concatenating, several fully-connected layers are employed to predict the input data, which forms the most important part of our proposed model. The layers include two hidden layers with tanh activation function and an output layer with linear activation function used as an ensemble predictor which are responsible for ensembling. In principle, the fully-connected layers form the MLP method that performs non-linear transformations. Finally, at the end of the network, prediction is done. The shrinkage phase, which is the most important part of our model, includes two regression methods: LASSO and EN.

TABLE 2. The parameter settings of MLP Ensemble model.

Layer	Configuration	Activation function	
Convolution 1	Filters = 64; kernel size = 2	tanh	Loss function = mse
Max-pooling 1	Kernel size = 2		Optimizer = adam
Convolution 2	Filters = 64; kernel size = 3	tanh	Batch size = 64
Max-pooling 2	Kernel size = 2		Epoch = 20
Fully connected	Units = 32	Relu	
Fully connected	Units = 1	Linear	
Concatenated layer			
Fully connected	Units = 32	tanh	
Fully connected	Units = 32	tanh	
Fully connected	Units = 1	Linear	

These methods are used to reduce the number of DRs and select the most suitable DR to improve the model performance. The combination of the mentioned methods is compared with the non-shrinkage method. Figure 4 shows the structure of the MLP Ensemble model.

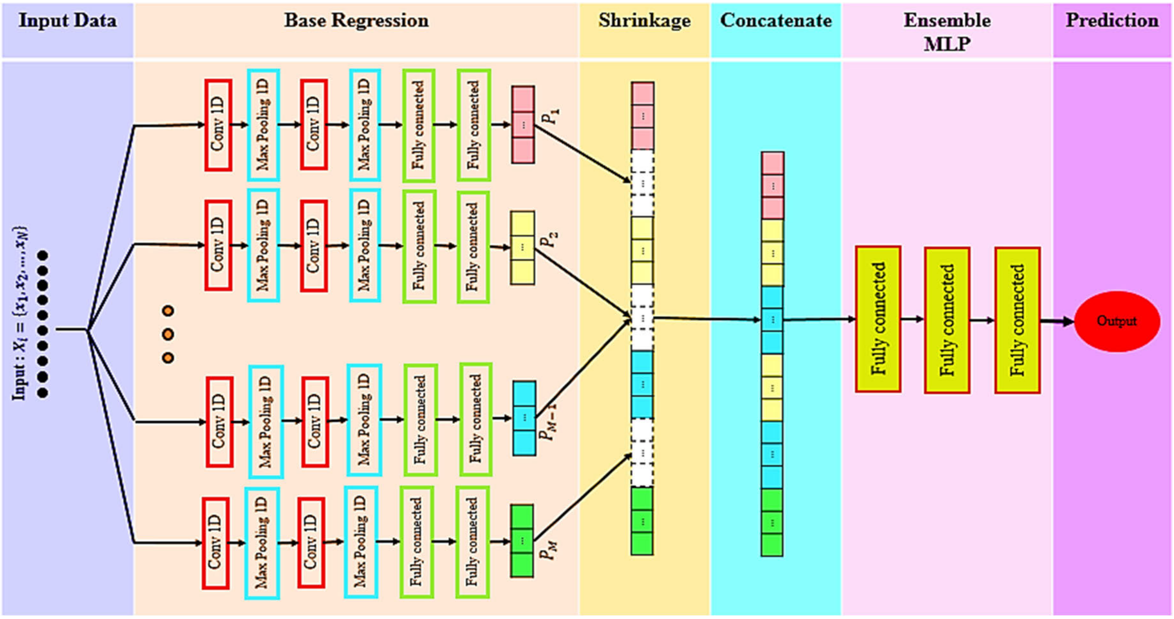


FIGURE 4. The structure of MLP nsemble model.

TABLE 3. The parameter settings of the weighted average ensemble model.

Layer	Configuration	Activation function	

Convolution 1	Filters = 64;
	kernel size = 2
		tanh	Loss function = mse
			Optimizer = adam
			Batch size = 64
			Epoch = 20

Max-pooling 1	Kernel size = 2		
Convolution 2	Filters = 64; kernel size = 3	tanh	
Max-pooling 2	Kernel size = 2		
Fully connected	Units = 32	Relu	
Fully connected	Units = 1	Linear	
Concatenated layer			
Fully connected	Units = 1	Linear	

TABLE 4. The parameter settings of the simple average ensemble model.

Layer	Configuration	Activation function	

Convolution 1	Filters = 64;
	kernel size = 2
		tanh	Loss function = mse
			Optimizer = adam
			Batch size = 64
			Epoch = 20

Max-pooling 1	Kernel size = 2		
Convolution 2	Filters = 64; kernel size = 3	tanh	
Max-pooling 2	Kernel size = 2		
Fully connected	Units = 32	Relu	
Fully connected	Units = 1	Linear	
Average Layer			

D. MODEL 2: WEIGHTED AVERAGE ENSEMBLE

Similar to the MLP ensemble model, described in subsection C (Model 1), this model also includes three phases: base regression, shrinkage, and ensemble. The base regression phase includes several parallel DRs. In the shrinkage phase, two methods, i.e. LASSO and EN, are used, which are responsible for selecting DRs. The trained DRs are ensembled by applying the shrinkage methods and without them. In this model, the ensemble method is WA, which is generated by combining the single fully-connected layer and linear activation function. In other words, the WA ensemble method performs a linear transformation. Finally, prediction is done at the end of the network. Figure 5 and Table 3 show the structure of the WA ensemble model and its parameter setting, respectively.

E. MODEL 3: SIMPLE AVERAGE ENSEMBLE

Similar to the two previous models, described in Subsections C (Model 1) and D (Model 2), the SA ensemble model is also constructed of similar phases. The difference between this model with other models is just the ensemble phase. For ensembling, the Simple Average method is used instead of MLP and WA. It should be noted that all models have shrinkage phases. In this phase, the number of DRs are reduced by LASSO and EN methods so that the selected effective DRs, which improve and increase the model accuracy, is remained in the model. Finally, these made models are compared with non-shrinkage model.

F. IMPLEMENTATION PLATFORM AND LIBRARIES

This experiment was performed on a system with an Intel[1] core[2] i7-6700K CPU @ 4.00 GHz accelerated by NVIDIA GeForce TITANX Graphics and 64 Gb memory. In the implementation process, a tremendous amount of parameters is created for which an ordinary CPU takes considerable execution time. To overcome this problem, a GPU accelerator is used to build the model to save a large amount of time. The in-depth learning approach, represented in our paper,

[1] Registered trademark.
[2] Trademarked.

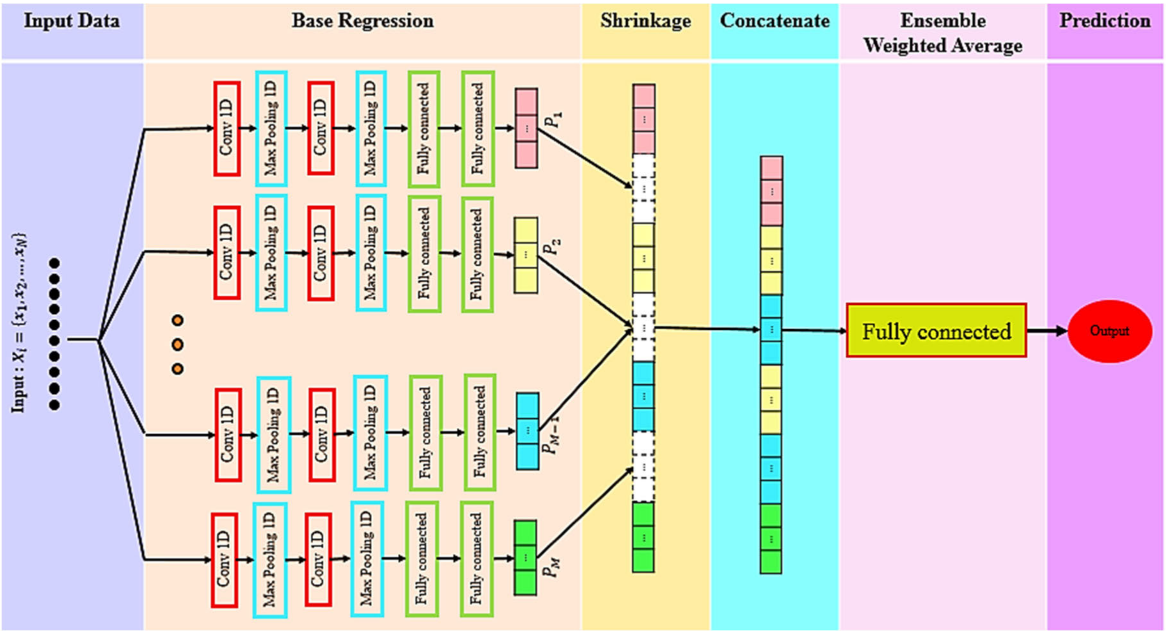


FIGURE 5. The structure of weighted average ensemble model.

is built using R programming language. The libraries used in our model are Keras, lattice, caret, and glmnet. The glmnet package is used for LASSO and EN and Keras package is used for CNN.

IV. EXPERIMENT RESULTS AND DISCUSSION

In this section, we will present the performance of the ERDeR model through a Monte Carlo simulation study and three real datasets including Boston House Price, Real Estate Valuation, and Gold Price Per Ounce. Then, their prediction performance is will be compared based on the number of DRs, shrinkage methods, and different ensemble approaches. The results will be presented in subsections B and C of this section. The ensemble approaches, which are divided into three approaches namely MLP, WA, and SA along with the shrinkage methods such as LASSO and EN, are compared with them. After combining, they produce new hybrid models that are compared with non-shrinkage models.

A. SIMULATION STUDY DESIGN

A Monte Carlo simulation study is conducted to assess the performance of the proposed model in R. This is done by applying shrinkage and non-shrinkage methods to improve the performance of ERDeR, which consists of three phases. The phases include base regression to train DRs as base learners, shrinkage to select the most suitable DRs and reduce the number of them, and ensemble methods to aggregate. Therefore, the proposed model is evaluated based on different criteria of MSE, RMSE, MAE, and R^2 as well as the number of selected DRs, type of shrinkage methods, and ensemble approaches.

In this study, we assume that the initial dataset is generated randomly from the standard normal distribution $N(0, 1)$ for the variables of $x_1, x_2, x_3, x_4, x_5, x_6, x_7$, which includes N = 500 observations with p = 7 predictions in the linear model. The following linear regression model are used to make the response variable:

$$\begin{aligned} y_1 &= 3 - 5.5x_1 + 7.3x_2 + 9x_3 + 10.3x_4 - 7x_5 \\ &\quad + 1.5x_6 + 0.9x_7 \\ y &= y_1 + \varepsilon \end{aligned} \tag{7}$$

where ε is the normal distribution $N(0,\sigma^2)$ and σ^2 is equal to $\frac{1}{3}$ of the standard deviation of y_1.

First, the standardization strategy is used for data preprocessing. The equation is as follows:

$$x_{stand} = \frac{x - mean\,(x)}{sd\,(x)} \tag{8}$$

where x represents the sample and mean(.) and sd(.) indicate the mean and standard deviation of x, respectively.

The simulation dataset is divided into training and test set. 80% of the data is assigned to the training set and the rest is dedicated to the test set. To train the network, we use deep CNN as base learners. The learners include 30, 40, and 50 parallel DRs that receive the input and provide the corresponding prediction. In the shrinkage and ensemble phases, the number of DRs are reduced and the remaining DRs are aggregated, respectively.

In the base regression phase, each of the DRs contains two convolution and max pooling layers. To do the convolution operation, the dataset is entered to the network of 30, 40, and 50 parallel DRs. In the convolution layer, 1D-CNN with two convolutions and max-pooling layers is used. The first and second convolution layers are set to 2 and 3 kernel sizes with 64 filters, 2 pooling sizes, and tanh activation function, respectively. Finally, two fully-connected layers with 32 units and ReLU activation function for the hidden layer and one unit and linear activation function for output layer are used. These features compute the prediction value of networks.

After training the DRs, in the fusion phase, concatenation is used to merge parallel DRs. In the ensemble phase, three methods including MLP, WA, and SA are applied to the ensemble. In the MLP method, three fully-connected layers are used for prediction and in its structure, two fully-connected layers with tanh activation function, 32 units in hidden layer and linear function and one unit are used in the output layer. Another method is WA in which a fully-connected layer with unit =1 and linear activation function is used to perform weighted averaging. Finally, the third method to the ensemble is SA, which employs simple averaging. The described methods are combined with shrinkage methods, which will be explained in continuation, and produce new models. They are compared with non-shrinkage models in which no shrinkage method is used. In the last phase of the proposed model, the LASSO and EN regressions are applied to shrink the DRs. These methods were described in Section II. Finally, some DRs are automatically selected without prior selection. Some others are removed and the remaining ones are aggregated.

The most important part of this research is the use of shrinkage methods. Applying these methods cause to reduce the number of DRs and improve the performance of the model. A linear regression model is defined on the output of parallel DRs, which is as follows:

$$\bar{Y} = C\beta + \varepsilon; \varepsilon \sim N(0, 0.5) \tag{9}$$

where $\bar{Y}$ is the average of DRs based on MLP, WA, and SA ensemble methods. C and β are the output of parallel DRs and linear regression coefficients, respectively. The β value for 50 DRs is equal to:

$$\beta = (\underbrace{1, 2, 3, 4, 5, 6, 7, \ldots, 1, 2, 3, 4, 5, 6, 7,}_{s=21} \underbrace{0, \ldots, 0}_{p-s=2})'$$

In the last phase of the proposed model, the LASSO and EN regressions are applied to shrink the DRs. These methods were described in Section II. Finally, some DRs are automatically selected without prior selection. Some others are removed and the remaining ones are aggregated.

One of the most important parts of DL algorithms is hyperparameters. Among these hyperparameters, it can be referred to epoch and batch size. Each epoch is defined as the number of iterations of the learning algorithm. The number of epochs can have different values. A large number of epochs lead to the increase in training time. Therefore, 20 epochs are assumed in the proposed model. Each batch size refers to the number of training samples utilized in one iteration. The batch size is 64. Other details of parameter settings about the proposed models are given in Tables 2, 3, and 4, respectively.

B. SIMULATION RESULTS

To evaluate the performance of MLP-NON, MLP-EN, and MLP-LASSO, based on the MLP ensemble method, WA-NON, WA-EN, and WA-LASSO based on the WA ensemble method, and SA-NON, SA-EN, and SA-LASSO based on SA ensemble method, we employ the simulation dataset to show how the proposed models perform and then compare them. The results of these models are calculated with 100 iterations. The comparison criteria include Mean Square Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of determination (R^2). They are defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{10}$$

$$MSE = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N} \tag{11}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}} \tag{12}$$

$$MAE = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N} \tag{13}$$

where $\hat{y}_i$, y_i, and N are the predicted values, true values of ith sample, and sample size, respectively.

In this case study, the MLP, WA, and SA methods in combination with shrinkage methods and non-shrinkage models are compared. Additionally, this comparison is done to demonstrate the appropriate performance of the shrinking methods and ensemble approaches in combination with deep CNNs that are used as a DR. As described in Section IV, 30, 40, and 50 parallel DRs are used to compare the combination of ensemble and shrinkage methods. The obtained results of the ERDeR framework based on the simulation dataset are shown in Tables 5-7. According to Tables 5 and 6, the MLP-EN and WA-EN achieve the best MSE, RMSE, MAE and R^2 values in all states of DRs. It can be said that increasing the number of DRs does not change the results but EN in combination with MLP and WA have better performance in all the three states. In other words, the obtained results confirm that MLP-EN and WA-EN gain the highest prediction; while in the MLP-LASSO and WA-LASSO models, the accuracy is slightly less than MLP-NON and WA-NON. Based on these results, if WA is used instead of MLP, WA can slightly achieve better performance. In other words, using one fully-connected layer for ensembling can have better performance than several fully-connected layers.

The WA ensemble method, which is combined with 50 parallel DRs and one fully-connected layer, makes a structure with less error, higher accuracy, and more suitable model than 30 and 40 parallel DRs. But the MLP method with 30 DRs and three fully-connected layers slightly increase the error of the ERDeR model, so the increase in the number of DRs increases its error as well. The variety of ensemble methods lead to various network architecture, as a result of which the performance of the network is changed. In the

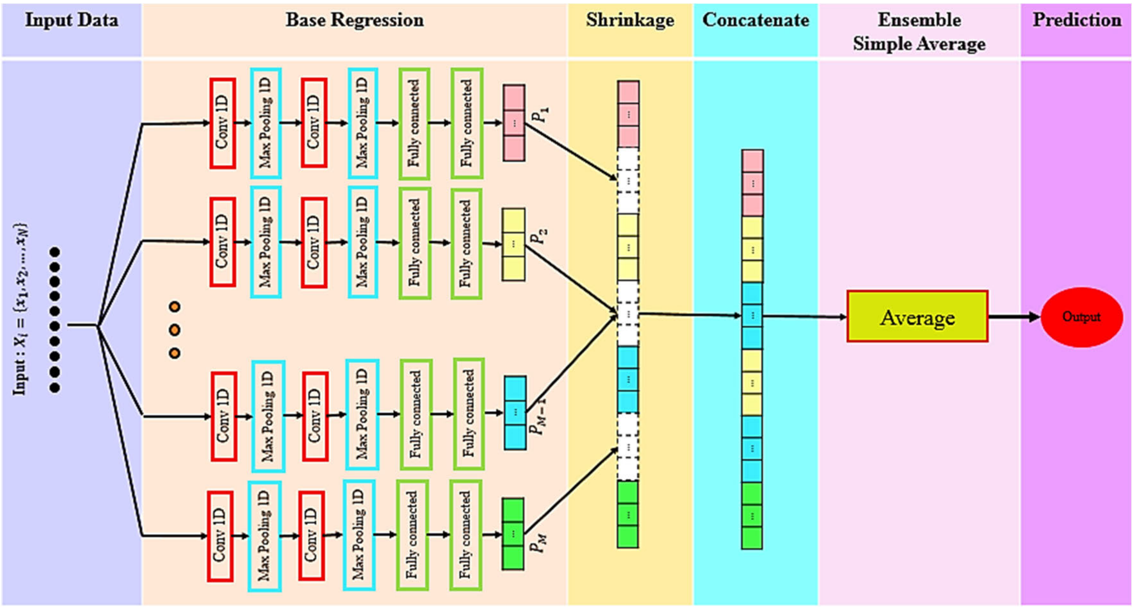
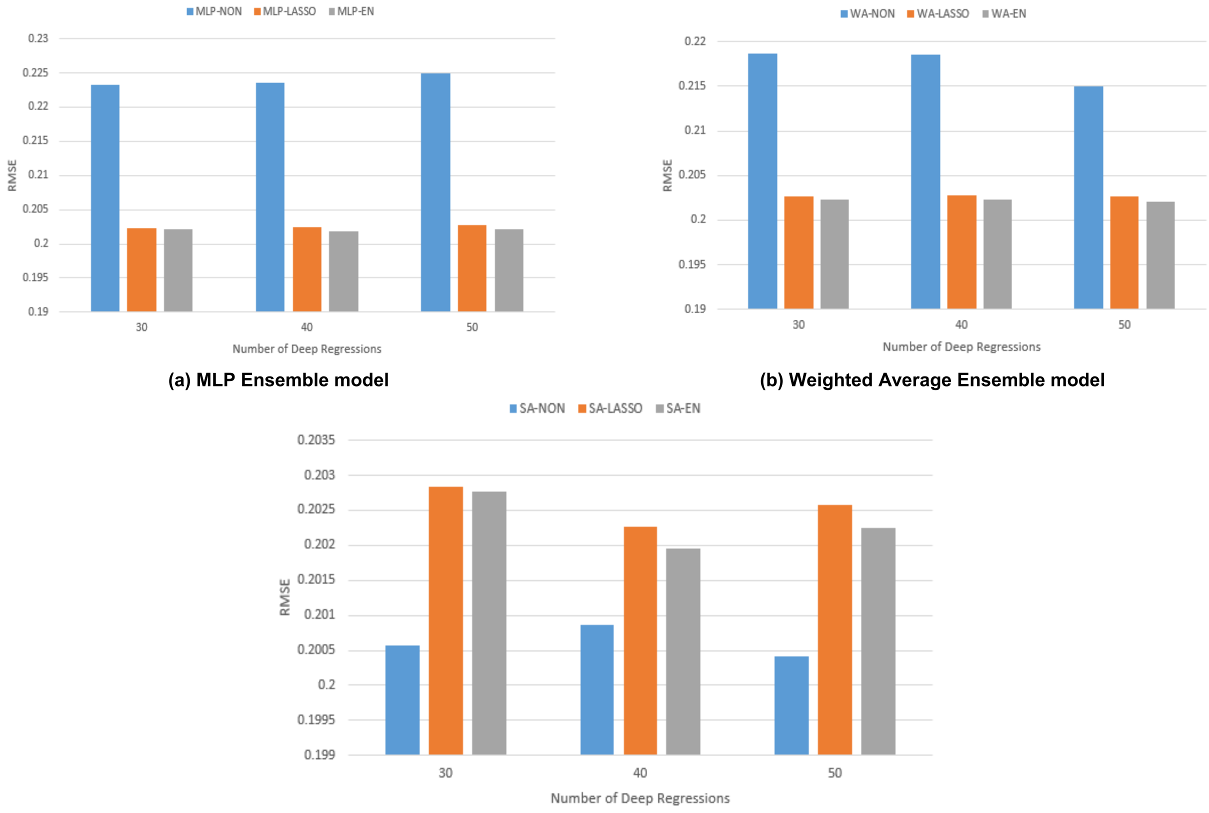


FIGURE 6. The structure of simple average ensemble model.



(a) MLP Ensemble model

(b) Weighted Average Ensemble model

(c) Simple Average Ensemble model

FIGURE 7. Bar-plot of comparison between three ensemble approaches and their combination with shrinkage methods on simulation dataset. Comparison of ERDeR is carried out with applying shrinkage methods and without them. (a) The RMSE value of MLP-EN and MLP-LASSO on different numbers of DRs. The MLP-EN is the lowest value and has better performance among three models. (b) The RMSE value of WA-EN and WA-LASSO on different numbers of DRs. The WA-EN is the lowest value and has better performance among three models. (c) The RMSE value of SA-EN and SA-LASSO on different numbers of DRs. The SA-NON is the lowest value and has better performance among three models.

architecture with one fully-connected layer that leads to making the WA ensemble model, the error rate is slightly less than three fully-connected layers that generat the MLP ensemble model.

TABLE 5. A comparison of ERDeR model results based on combining shrinkage methods and MLP Ensemble approach on simulation dataset.

Number of DRs	Algorithm	MSE	RMSE	MAE	R^2
30	MLP-NON	0.050259387	0.22336114	0.175526158	0.949651178
	MLP-LASSO	0.041108258	0.202368041	0.159807345	0.958568035
	MLP-EN	**0.041036576**	**0.202194566**	**0.159683363**	**0.958639609**
40	MLP-NON	0.05031539	0.223641598	0.176224859	0.949526212
	MLP-LASSO	0.041135899	0.202403601	0.159885932	0.958547165
	MLP-EN	**0.04095105**	**0.201966407**	**0.159531466**	**0.958740538**
50	MLP-NON	0.050856618	0.224905062	0.177061515	0.949084531
	MLP-LASSO	0.041285366	0.202808281	0.160165092	0.958446886
	MLP-EN	**0.041008977**	**0.202115575**	**0.159679494**	**0.958728197**

TABLE 6. A comparison of ERDeR model results based on combining shrinkage methods and weighted average ensemble approach on simulation dataset.

Number of DRs	Algorithm	MSE	RMSE	MAE	R^2
30	WA-NON	0.04815965	0.218638441	0.173372732	0.951461182
	WA-LASSO	0.041249178	0.202704165	0.160174631	0.958441641
	WA-EN	**0.041093694**	**0.202327214**	**0.15986679**	**0.958592188**
40	WA-NON	0.048102497	0.218593014	0.17278633	0.951510762
	WA-LASSO	0.041269092	0.202747166	0.160276812	0.958428161
	WA-EN	**0.041088609**	**0.202305653**	**0.159843002**	**0.95860871**
50	MLP-NON	0.046458204	0.215012118	0.170661394	0.953122389
	WA-LASSO	0.041244889	0.2027071	0.160090258	0.958435067
	WA-EN	**0.040998121**	**0.202105574**	**0.159474756**	**0.958672282**

TABLE 7. A comparison of ERDeR model results based on combining shrinkage methods and simple average ensemble approach on simulation dataset.

Number of DRs	Algorithm	MSE	RMSE	MAE	R^2
30	**SA-NON**	**0.040404693**	**0.20057658**	**0.159499111**	**0.959234104**
	SA-LASSO	0.041291604	0.202842477	0.160099839	0.958383451
	SA-EN	0.04126596	0.202766929	0.160019105	0.958405659
40	**SA-NON**	**0.040521146**	**0.200872354**	**0.159812285**	**0.959149098**
	SA-LASSO	0.041061547	0.202260215	0.159721183	0.958613742
	SA-EN	0.04093824	0.201956779	0.159519039	0.958743105
50	**SA-NON**	**0.040357059**	**0.200409111**	**0.15971765**	**0.959349694**
	SA-LASSO	0.041200306	0.202585816	0.16000727	0.958462831
	SA-EN	0.041064241	0.202256231	0.159760234	0.958605504

Table 7 shows the obtained results based on combining shrinkage methods and SA ensemble approach. In this model, the SA-NON model outperforms the other two models, i.e. SA-LASSO and SA-EN. The main reason for this issue is the use of fully-connected layers in MLP and WA models where weights are applied to the generated outputs by the CNN. These weights construct non-linear combinations of features and consequently generate non-linear probabilities of predictions. As can be seen in Table 7, the MSE of 50 parallel DRs in SA-NON, SA-LASSO, and SA-EN models are 0.04035, 0.04120, and 0.04106, respectively. In 30 and 40 parallel DRs, this order is also established which shows that the error of SA-NON is less than the two other combined methods in all cases and it has better performance. This means that increasing the number of DRs does not make change in the performance of SA-NON model.

The bar plots of the obtained results for different models in terms of RMSE criterion are given in Figure 7. As can be

TABLE 8. A comparison of ERDeR model results based on combining shrinkage methods and MLP ensemble approach on boston house price dataset.

Number of DRs	Algorithm	MSE	RMSE	MAE	R-square	Number of selected DRs
30	MLP-NON	0.206191703	0.449184323	0.312329668	0.797629619	30
	MLP-LASSO	0.19242716	0.434704393	0.278005945	0.815384704	29
	MLP-EN	**0.192325017**	**0.433968769**	**0.274100341**	**0.81573632**	**28**
40	MLP-NON	0.175274715	0.418658232	0.310127388	0.829124582	40
	MLP-LASSO	0.14924516	0.386322612	0.273950814	0.854500809	24
	MLP-EN	**0.147829925**	**0.384486574**	**0.273101489**	**0.855880523**	**23**
50	MLP-NON	0.187153381	0.432612276	0.30700688	0.817544063	50
	MLP-LASSO	0.154828172	0.393482112	0.279302147	0.849057928	18
	MLP-EN	**0.150500927**	**0.387944489**	**0.275021291**	**0.853276562**	**21**

seen, applying shrinkage methods on MLP and WA ensemble methods outperform the SA; while among the used shrinkage methods, the EN performs better than LASSO. Further details in Figure 7b indicate that the RMSE of 30 parallel DRs is 0.20232 in the WA-EN model and 0.2027 in the WA-LASSO model. Moreover, in 40 and 50 parallel DRs, the RMSE is equal to 0.20231 and 0.2021 in the WA-EN and 0.20274 and 0.20271 in the WA-LASSO. Therefore, it can be concluded that the WA-EN model has less error and better performance. In the SA, the RMSE of 30 parallel DRs is equal to 0.2005 in the SA-NON model and 0.2028 and 0.2027 in the SA-LASSO and SA-EN, respectively.

In addition, the RMSE of 40 and 50 parallel DRs is equal to 0.2008 and 0.2004 in SA-NON model, is equal to 0.20226 and 0.2025 in SA-LASSO, and 0.2019 and 0.20225 in SA-EN model, respectively. Consequently, it can be inferred that the SA-NON model has better performance. This shows that the ERDeR prediction model partly depends on the type of ensemble methods. Besides, the performance of shrinking methods depends on the type of methods, i.e. LASSO and EN. Overall, the MLP and WA approaches in combination with shrinkage methods and comparison with non-shrinkage models show that automatically removing the number of DRs increases the ERDeR accuracy.

C. REAL DATA ANALYSIS

In this section, the performance of the ERDeR model by combining different shrinkage methods, and ensemble approaches are described on three real datasets. The obtained results will be given in three parts.

1) CASE STUDY 1: EXPERIMENT RESULTS ON BOSTON HOUSE PRICE DATASET

The Boston House Price dataset is used to study the performance of the proposed models. This dataset contains 506 observations and 13 variables whose response variable is the median of the owner-occupied house price. The data were first published by Harrison and Rubinfeld [89], which is publicly available through the MASS package in R (https://cran.r-project.org/package=MASS). Like simulation section, first, the data are standardized using Equation (8). Then, they are divided into training and test sets. 80% and 20% of the intended dataset are considered as training and test sets, respectively. The training and test sets are selected randomly by the network. The training set is used as a subset of the initial observations to train the prediction models, while the test set is used to validate and evaluate the model accuracy.

Tables 8 and 9 indicate the great performance of combining shrinking methods with MLP and WA ensemble methods. 1D-CNN is used as a deep regression. To reduce the number of DRs, the LASSO and EN are applied. These methods eliminate ineffective DRs of the model by applying a penalty on the sum of squared errors and create the most suitable model by reducing their number. This new created model will be accompanied by increasing accuracy and improvement. Therefore, combining these methods with ensemble approaches, i.e. MLP and WA, can increase accuracy and improve the performance of the proposed new model. Based on the presented results in Tables 8 and 9, it can be seen that the best prediction is related to the MLP-EN and WA-EN models with the least error in terms of MSE, RMSE, MAE, and R^2. The MLP method with three fully-connected layers in combination with EN has the highest R^2 and the lowest MSE, RMSE, and MAE values than MLP-NON and MLP-LASSO, respectively.

Moreover, it selects 28, 23, and 21 parallel DRs from 30, 40, and 50 parallel DRs, respectively. The MSE of MLP-EN is equal to 0.1923, 0.1478, and 0.1505 in 30, 40, and 50 parallel DRs, respectively. Changing the ensemble method and using WA can have better performance than MLP, which can be seen in its result in Table 9. In this method, the WA-EN model with one fully-connected layer by selecting 22, 20, and 36 parallel DRs from 30, 40, and 50 parallel DRs have better performance, respectively. By contrast, using SA changes the performance of the model. Unlike the two other ensemble approaches, this one does not provide excellent performance in combination with shrinkage methods, which can be due to

TABLE 9. A comparison of ERDeR results based on combining shrinkage methods and weighted average ensemble approach on boston house price dataset.

Number of DRs	Algorithm	MSE	RMSE	MAE	R-square	Number of selected DRs
30	WA-NON	0.204500662	0.452217494	0.315589184	0.80063219	30
	WA-LASSO	0.144961032	0.380737485	0.271730155	0.858677408	19
	WA-EN	**0.144849561**	**0.380591068**	**0.270848643**	**0.858786082**	**22**
40	WA-NON	0.180745329	0.42514154	0.30654984	0.823791277	40
	WA-LASSO	0.148489589	0.385343468	0.273187236	0.855237416	25
	WA-EN	**0.14615374**	**0.382300589**	**0.272034687**	**0.857514637**	**20**
50	WA-NON	0.156496559	0.395596459	0.286224576	0.847431417	50
	WA-LASSO	0.142300737	0.377227699	0.265357866	0.861270931	44
	WA-EN	**0.141523027**	**0.376195463**	**0.264175885**	**0.862029122**	**36**

TABLE 10. A comparison of ERDeR model results based on combining shrinkage methods and simple average ensemble approach on boston house price dataset.

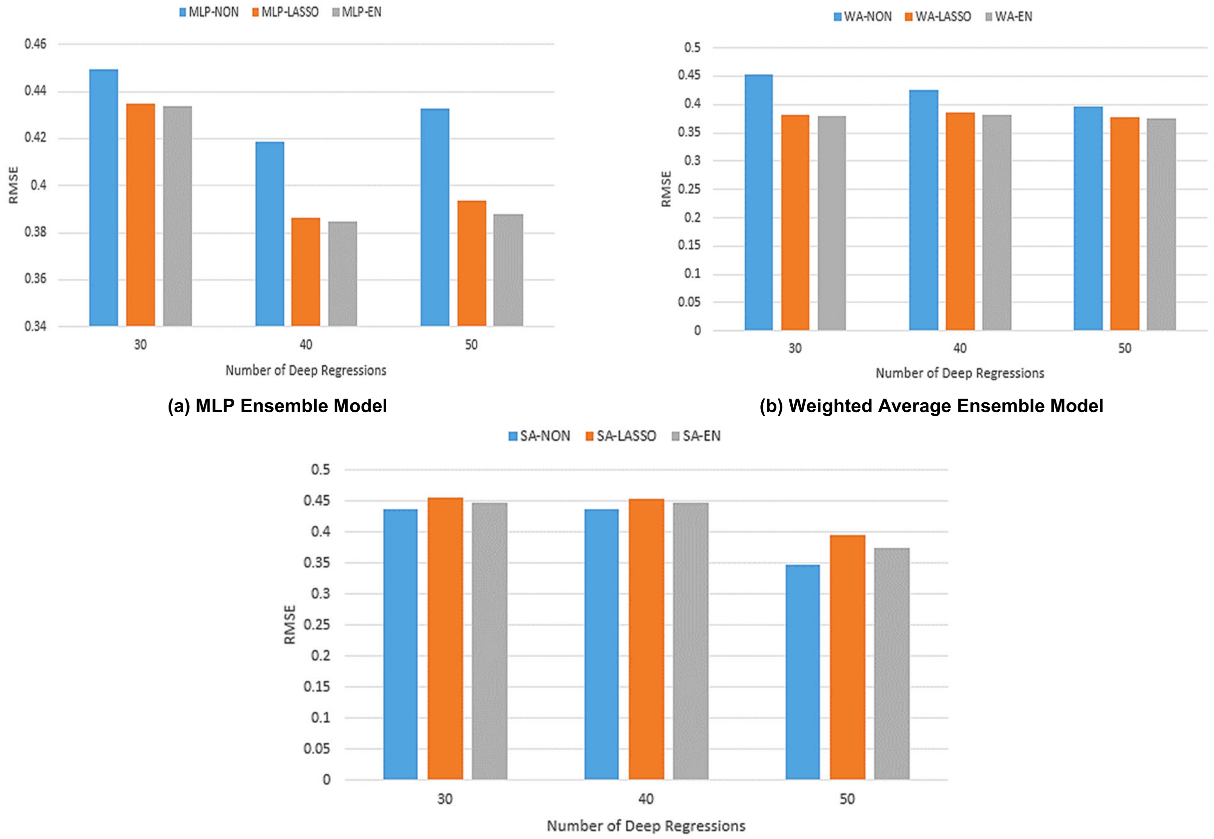
Number of DRs	Algorithm	MSE	RMSE	MAE	R-square	Number of selected DRs
30	**SA-NON**	**0.194442009**	**0.436599393**	**0.811008646**	**0.30506282**	**30**
	SA-LASSO	0.211768019	0.455147645	0.79790545	0.29022654	13
	SA-EN	0.20436775	0.447697229	0.804267656	0.28304546	15
40	**SA-NON**	**0.193373262**	**0.436241572**	**0.298995112**	**0.81344332**	**40**
	SA-LASSO	0.209475032	0.452974488	0.292691277	0.799193005	24
	SA-EN	0.204256071	0.44746371	0.288173846	0.803991602	27
50	**SA-NON**	**0.120580893**	**0.347247596**	**0.25704749**	**0.88244562**	**50**
	SA-LASSO	0.156619514	0.395751833	0.277322774	0.847311548	16
	SA-EN	0.140559026	0.374912025	0.264751908	0.862968926	45

the fully-connected layers and the non-linear possibilities in the MLP.

According to the obtained results from this method, which can be seen in Table 10, the MSE of SA-NON is equal to 0.1944, 0.1933, and 0.1205 for 30, 40, and 50 parallel DRs, respectively. Therefore, it has better performance than SA-LASSO and SA-EN. Figure 8 illustrates the bar-plot of applied different methods on the Boston house price dataset. These diagrams represent the performance of the ERDeR model based on the number of DRs, shrinkage methods and ensemble approaches. As can be seen in Figures 8a and 8b, combining shrinkage methods with MLP and WA ensemble approaches significantly outperform significantly the SA. Based on these plots, it can be stated that if the WA and MLP are exploited as ensemble methods, the shrinkage methods such as WA-LASSO, WA-EN, MLP-EN, and MLP-LASSO will have better performance than WA-NON and MLP-NON. Figure 8c demonstrates the SA results indicating the better performance of SA-NON. Further details of the SA indicate that the RMSE of 30 parallel DRs is equal to 0.4365 in the SA-NON and is 0.4476 in the SA-LASSO model. In addition, the RMSE of 40 and 50 parallel DRs which is 0.4362 and 0.3472 in SA-NON has the best performance. Also, the RMSE of 40 and 50 parallel DRs which is equal 0.4529 and 0.3957 in the SA-LASSO has the worst performance. Therefore, it can be concluded that the SA-NON model has less error and outperforms SA-LASSO and SA-EN.

2) CASE STUDY 2: EXPERIMENT RESULTS ON REAL ESTATE VALUATION DATASET

To study the performance of the ERDeR model, the Real Estate Valuation dataset is used. This dataset is from the UCI machine learning repository and the original owner of this dataset is Yeh and Hsu [32]. It consists of seven features from which five features are selected including house age, distance to the nearest MRT station, number of convenience stores in the living circle on foot, latitude, longitude, the transaction date, and house price of the select unit area as the independent variable and house price of the unit area as the dependent variable. Similar to the Boston House Price dataset, the data are standardized using Equation (8). Then, it is separated into training and test set. The sample size in the training set and test sets is 332 (80%) and 82 (20%), respectively.



(a) MLP Ensemble Model

(b) Weighted Average Ensemble Model

(c) Simple Average Ensemble Model

FIGURE 8. Bar-plot of comparison between three ensemble approaches and their combination with shrinkage methods on Boston House Price dataset. Comparison of ERDeR is carried out with applying shrinkage methods and without them. (a) The RMSE value of MLP-EN and MLP-LASSO on different numbers of DRs. The MLP-EN is the lowest value and has better performance among three models. (b) The RMSE value of WA-EN and WA-LASSO on different numbers of DRs. The WA-EN is the lowest value and has better performance among three models. (c) The RMSE value of SA-EN and SA-LASSO on different numbers of DRs. The SA-NON is the lowest value and has better performance among three models.

TABLE 11. A comparison of ERDeR model results based on combining shrinkage methods and MLP ensemble approach on real estate valuation dataset.

Number of DRs	Algorithm	MSE	RMSE	MAE	R-square	Number of selected DRs
30	MLP-NON	0.25600681	0.505971155	0.389597182	0.68212326	30
	MLP-LASSO	0.227897377	0.477385984	0.366499146	0.717025983	7
	MLP-EN	**0.221347568**	**0.470475896**	**0.3601908**	**0.725158704**	**16**
40	MLP-NON	0.243995874	0.493959385	0.374465205	0.697036915	40
	MLP-LASSO	0.227405388	0.476870411	0.366103961	0.717636873	38
	MLP-EN	**0.226575307**	**0.475999272**	**0.364886608**	**0.718667562**	**35**
50	MLP-NON	0.28405678	0.532969774	0.420159325	0.64729437	50
	MLP-LASSO	0.222163633	0.471342374	0.361468041	0.724145418	37
	MLP-EN	**0.221604801**	**0.470749192**	**0.36078939**	**0.724839305**	**38**

Experiments indicate that the type of shrinkage methods and ensemble approaches has a significant impact on model performance. Tables 11-13 show the obtained results from the Real Estate Valuation dataset by different ensemble approaches. Tables 11 and 12 show the performance of the model in combination with ensemble approaches, i.e. MLP, WA, and shrinkage methods, i.e. LASSO and EN. To reduce the number of parallel DRs and improve the model performance, the shrinkage methods are applied on base learners. Additionally, as one of the ensemble methods, the WA method has the best performance. Table 13 illustrates the performance of the proposed model based on the SA

TABLE 12. **A comparison of ERDeR model results based on combining shrinkage methods and weighted average ensemble approach on real estate valuation dataset.**

Number of DRs	Algorithm	MSE	RMSE	MAE	R-square	Number of selected DRs
30	WA-NON	0.238199371	0.488056729	0.389458639	0.704234277	30
	WA-LASSO	0.227759365	0.477241412	0.367712926	0.717197349	20
	WA-EN	**0.227441072**	**0.476907824**	**0.36747779**	**0.717592565**	**22**
40	WA-NON	0.272672251	0.522180286	0.411375842	0.661430232	40
	WA-LASSO	0.227160502	0.476613578	0.367391213	0.717940941	36
	WA-EN	**0.226911813**	**0.476352614**	**0.366598856**	**0.718249732**	**33**
50	WA-NON	0.355110584	0.595911557	0.447863961	0.559068782	50
	WA-LASSO	0.224067617	0.473357811	0.362465113	0.721781293	25
	WA-EN	**0.223559101**	**0.472820368**	**0.3620478**	**0.722412704**	**31**

TABLE 13. **A comparison of ERDeR model results based on combining shrinkage methods and simple average ensemble approach on real estate valuation dataset.**

Number of DRs	Algorithm	MSE	RMSE	MAE	R-square	Number of selected DRs
30	**SA-NON**	**0.215209816**	**0.46390712**	**0.361251636**	**0.732779786**	**30**
	SA-LASSO	0.230706041	0.480318687	0.369924721	0.713538541	6
	SA-EN	0.221082605	0.47019422	0.358483541	0.725487702	10
40	**SA-NON**	**0.222320706**	**0.471509**	**0.367011627**	**0.723950386**	**40**
	SA-LASSO	0.269481798	0.519116	0.367011627	0.665391732	6
	SA-EN	0.241391441	0.491316	0.367011627	0.700270769	8
50	**SA-NON**	**0.214130355**	**0.462742213**	**0.359331356**	**0.734120123**	**50**
	SA-LASSO	0.223748377	0.473020483	0.366689425	0.722177684	7
	SA-EN	0.22213731	0.471314449	0.367311521	0.724178104	9

ensemble method. The performance of this model does not improve by applying the shrinkage methods while the non-shrinkage method shows better performance. As mentioned in the previous sections, selecting the suitable ensemble method can be an important factor in improving the model.

Various ensemble methods such as MLP, WA, and SA are used to optimize and improve the model. Their performance is displayed in Figures 9a, 9b, and 9c. Like the previous dataset, LASSO and EN shrinkage methods in combination with WA and MLP ensemble methods show the best performance due to reducing and selecting appropriate the number of DRs; while the SA method has an unsuitable performance. As seen in Figure 9c, despite the use of shrinkage methods, the SA cannot improve the model by reducing the number of DRs. However, other ensemble methods have better performance. After applying shrinkage methods, the WA approach makes a better model than MLP. It means that MLP-EN and WA-EN models have the best performance in comparison with SA-EN.

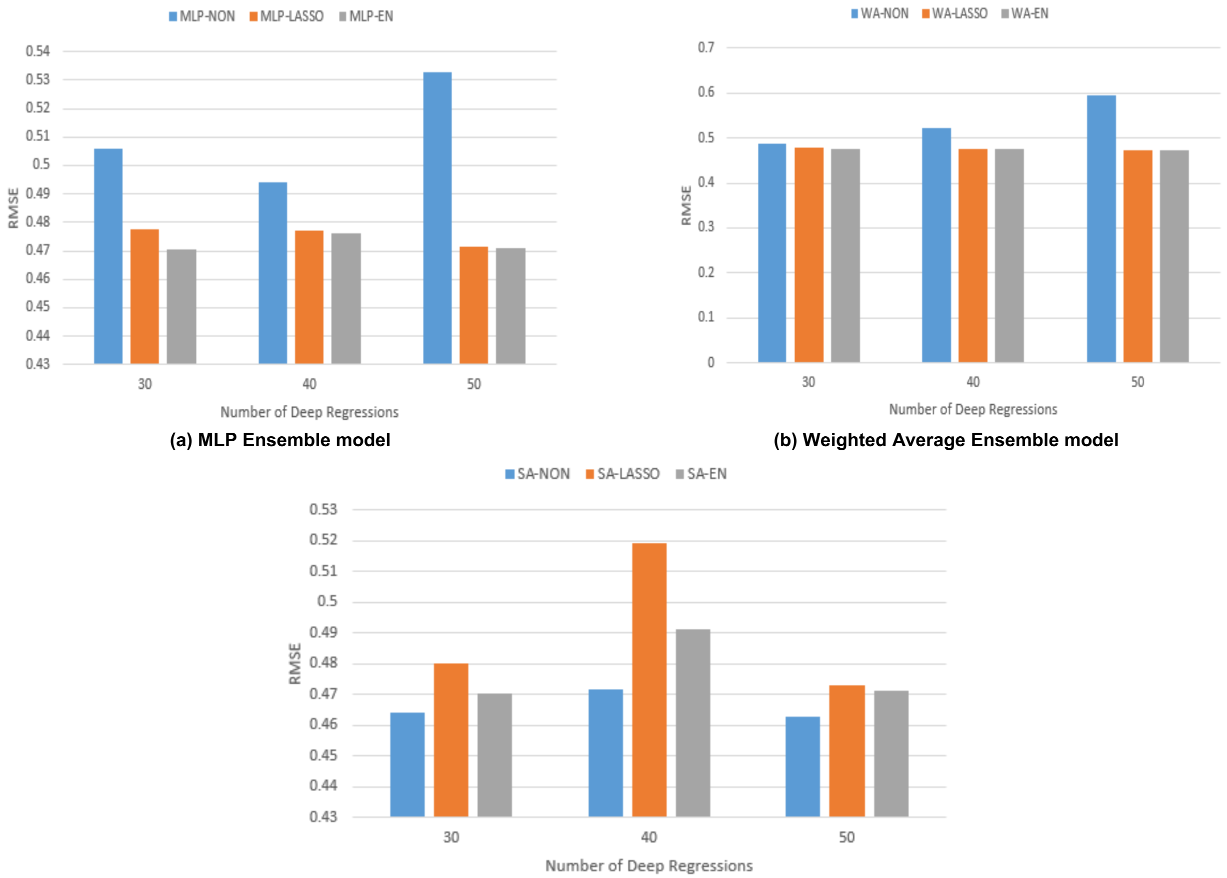
3) CASE STUDY 3: EXPERIMENT RESULTS ON GOLD PRICE PER OUNCE DATASET

In the third experiment, the Gold Price Per Ounce dataset is used to describe the performance of the proposed models. This data is publicly available on GitHub (https://github.com/cominsys/Data_GPPO_01), which consists of four features including high price, low price, open price, close price, and 717 observations. The dataset is time series data to convert it into a supervised learning problem, the sliding window [90] method is used. This technique takes as the following steps:

1) Determine the window size by length k.
2) Slid the window and shift forward one unit.
3) Continue the sliding until all windows are calculated.

In this research, we set the window size of length 10. Eventually, 40 independent variables and 707 observations were produced. Similar to the two previous datasets, the preprocessing was accomplished to standardize data. Then, it is separated into training and test set. The sample size in the training set and test set is 567 (80%) and 140 (20%), respectively.

In this proposed model, the effectiveness of shrinkage methods such as LASSO and EN in combination with ensemble approaches such as MLP, WA, and SA is considered. The least value is related to WA-EN and MLP-EN models, while MLP-NON and WA-NON are the highest error values. By contrast, the least error value belongs to SA-NON in



(a) MLP Ensemble model

(b) Weighted Average Ensemble model

(c) Simple Average Ensemble model

FIGURE 9. Bar-plot of comparison between three ensemble approaches and their combination with shrinkage methods on Real Estate Valuation dataset. Comparison of ERDeR is carried out with applying shrinkage methods and without them. (a) The RMSE value of MLP-EN and MLP-LASSO on different numbers of DRs. The MLP-EN is the lowest value and has better performance among three models. (b) The RMSE value of WA-EN and WA-LASSO on different numbers of DRs. The WA-EN is the lowest value and has better performance among three models. (c) The RMSE value of SA-EN and SA-LASSO on different numbers of DRs. The SA-NON is the lowest value and has better performance among three models.

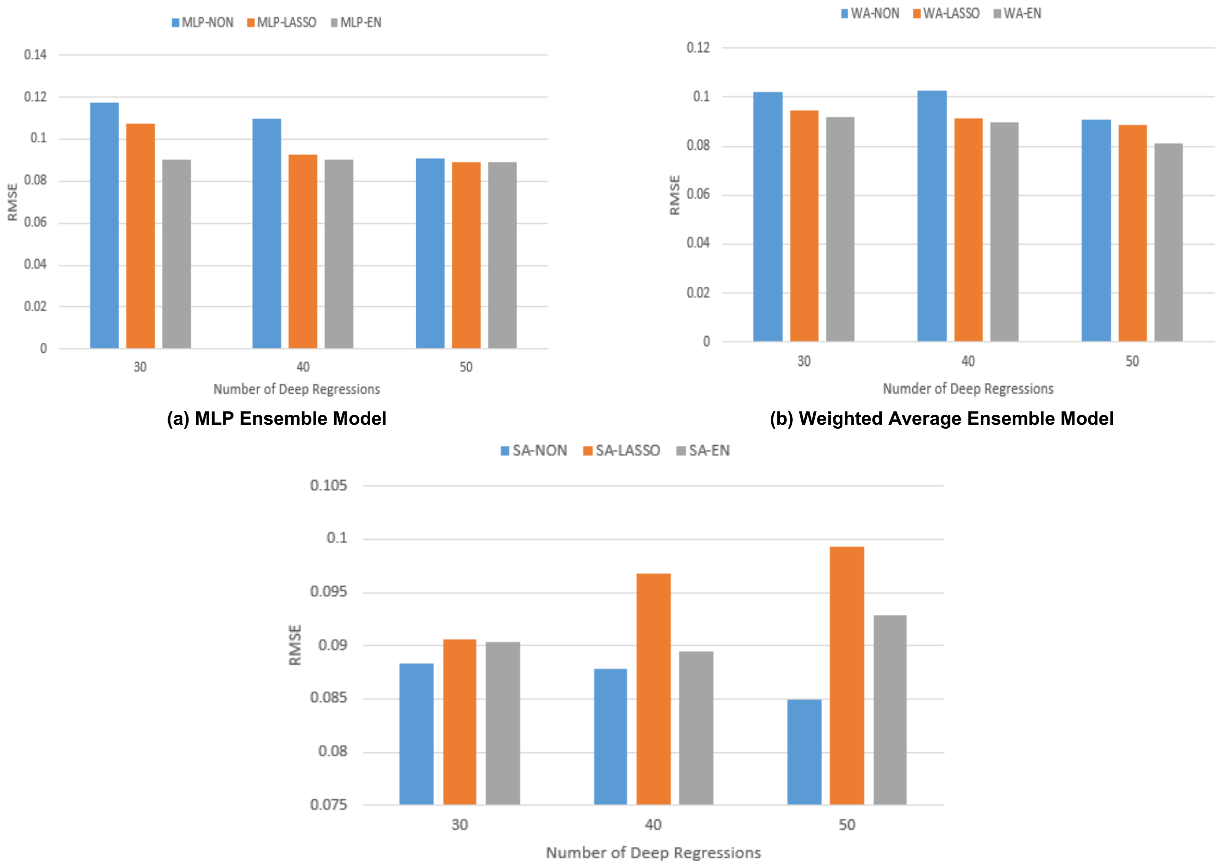
the SA. As a result, the effectiveness of shrinkage methods in combination with MLP and WA models is higher than other models. Additionally, we compare different shrinkage methods including LASSO, EN, and compare them with the non-shrinkage models based on different DRs. The results show that amongst the shrinkage methods, the EN has the best performance. More details are given in Tables 14-16.

The accuracy of MLP and WA approaches in combination with shrinkage methods and without them are illustrated in Figures 10a and 10b.

The results of the MLP method show that the RMSE after applying shrinkage methods on 30, 40, and 50 parallel DRs reach from 0.013, 0.012, and 0.0082 in MLP-NON to 0.00818, 0.00815, and 0.0078 in MLP-EN, respectively. In the WA method, this reduction is maintained. In the SA, applying the shrinkage methods not only do not improve the model, but also increase the RMSE from 0.0078, 0.0077, and 0.0072 in SA-NON to 0.0082, 0.0093, and 0.0098 in SA-LASSO.

D. COMPARISON WITH PAST STUDIES

In ensemble models, the performance of the model mainly depends on learners. Learners can be various algorithms, for example, in RF, Bagging, Boosting, and XGboost, the Decision Tree is used as learner [87]. In this paper, we proposed a new deep neural network that is used CNN algorithm as learners to increase the prediction accuracy by extracting more relationships of features. Currently, DL methods, particularly CNN architectures, have shown remarkable success in prediction. One of the notable advantages of CNNs is their ability to provide greater precision and improve system performance due to unique features, such as local connectivity and shared weights. Therefore, we decided to use CNN as a learner in our proposed model. Another factor that can address the model's complexity and computational load problem is using shrinkage methods to reduce the number of learners in ensemble models. These methods solve the overfitting problem by removing ineffective learners from the model. The number of learners and their type can affect prediction. Therefore, we attempted to improve prediction by replacing CNN in



(a) MLP Ensemble Model

(b) Weighted Average Ensemble Model

(c) Simple Average Ensemble Model

FIGURE 10. Bar-plot of comparison between three ensemble approaches and their combination with shrinkage methods on Gold Price per Ounce dataset. Comparison of ERDeR is carried out with applying shrinkage methods and without them. (a) The RMSE value of MLP-EN and MLP-LASSO in different numbers of DRs. The MLP-EN is the lowest value and has better performance among three models. (b) The RMSE value of WA-EN and WA-LASSO on different numbers of DRs. The WA-EN is the lowest value and has better performance among three models. (c) The RMSE value of SA-EN and SA-LASSO on different numbers of DRs. The SA-NON is the lowest value and has better performance among three models.

TABLE 14. A comparison of ERDeR model results based on combining shrinkage methods and MLP ensemble approach on gold price per ounce dataset.

Number of DRs	Algorithm	MSE	RMSE	MAE	R-square	Number of selected DRs
30	MLP-NON	0.013830504	0.117603165	0.091286965	0.985760886	30
	MLP-LASSO	0.01153871	0.107418388	0.080940004	0.988120389	8
	MLP-EN	**0.008182572**	**0.090457569**	**0.059736575**	**0.991575682**	**19**
40	MLP-NON	0.012057067	0.109804676	0.085135734	0.987586718	40
	MLP-LASSO	0.008527644	0.092345246	0.059922566	0.991220414	5
	MLP-EN	**0.008149485**	**0.090274499**	**0.05918757**	**0.991609746**	**16**
50	MLP-NON	0.008229563	0.09071694	0.068934936	0.991527302	50
	MLP-LASSO	0.00795781	0.089206556	0.058475952	0.991807084	4
	MLP-EN	**0.007880508**	**0.088772227**	**0.059232404**	**0.991886669**	**10**

the designed ERDeR model with DT in RF. This results in a model with higher precision compared to previous models.

Tables 17 and 18 explore a comparative study between the present study and other past proposed models. As can be seen, our proposed model (ERDeR) works well compared to the previous works in this field. In original RF algorithm, as can be seen in Tables 17 and 18, the MSE value of RF is larger than the other combined proposed models. One of the reasons for this issue might be the type of learners. The other reason can be the numbers of learners. To reduce the

TABLE 15. A comparison of ERDeR model results based on combining shrinkage methods and weighted average ensemble approach on gold price per once dataset.

Number of DRs	Algorithm	MSE	RMSE	MAE	R-square	Number of selected DRs
30	WA-NON	0.01045541	0.102251702	0.071304456	0.989235694	30
	WA-LASSO	0.008927278	0.094484274	0.062830103	0.990808974	6
	WA-EN	**0.008419567**	**0.091758197**	**0.060326489**	**0.991331685**	**16**
40	WA-NON	0.010474959	0.102347247	0.072718156	0.989215568	40
	WA-LASSO	0.00832558	0.091244615	0.060316604	0.991428449	7
	WA-EN	**0.008031426**	**0.089618221**	**0.058955719**	**0.991731293**	**18**
50	WA-NON	0.008212243	0.090621426	0.060231794	0.991545134	50
	WA-LASSO	0.00789225	0.088838337	0.061440422	0.991874581	3
	WA-EN	**0.006560135**	**0.08099466**	**0.055034876**	**0.993246052**	**16**

TABLE 16. A comparison of ERDeR model results based on combining shrinkage methods and simple average ensemble approach on gold price per once dataset.

Number of DRs	Algorithm	MSE	RMSE	MAE	R-square	Number of selected DRs
30	**SA-NON**	**0.007802201**	**0.088330067**	**0.058829962**	**0.99196729**	**30**
	SA-LASSO	0.008212839	0.090624717	0.060400602	0.99154452	7
	SA-EN	0.00815973	0.090331225	0.059955219	0.991599198	25
40	**SA-NON**	**0.007702207**	**0.087762221**	**0.05789033**	**0.992070238**	**40**
	SA-LASSO	0.009363088	0.096763049	0.067155376	0.990360288	7
	SA-EN	0.008014491	0.089523687	0.059257962	0.991748729	29
50	**SA-NON**	**0.007214017**	**0.084935368**	**0.056682651**	**0.992572852**	**50**
	SA-LASSO	0.009846857	0.099231329	0.072738532	0.989862227	6
	SA-EN	0.008624971	0.092870722	0.064010874	0.98938449	29

error, Wang and Wang [64] used a LASSO regression for tree selection of RF to automatically reduce the number of trees and control overfitting. Boston Housing Price dataset and Real Estate Valuation were used for training and testing, achieving RMSE of 3.13296 in Boston Housing Price dataset and 10.2755 in Real Estate Valuation.

The authors of [4] introduced a novel hybrid model called RARTEN that combined EN and RF. The results revealed a prediction with MSE of 9.7350 and MAE of 2.0906 on Boston Housing Price dataset. Experimental results and other penalized methods such as LASSO, Group Lasso, Adaptive Lasso demonstrated that RARTEN exhibits acceptable performance.

The results of a work called ECAPRAF [5] based on the combination of K-means clustering, RF, and penalized methods on Boston Housing Price dataset and Real Estate Valuation has 3.2218 and 5.5078 error value, respectively. The use of K-means to identify homogeneous subset of data resulted in a substantial improvement in the obtained results.

In this study, by changing learners from decision tree to deep regressions and replacing 1D-CNN instead of RF, a novel deep network was proposed. Additionally, different structures of ensemble leaning methods like MLP and WA were used instead of SA. By analyzing various metrics, it was concluded that the proposed work shows better results than other works in the literature. We used the same data for training and testing and achieved MSE, RMSE and MAE values of 0.1505, 0.3879, 0.2750 for ERDeR-MLP-EN, respectively. Comparing ERDeR-MLP-EN, ERDeR-WA-EN, and

TABLE 17. Comparison of existing works with our proposed model on boston housing price dataset.

Algorithm	MSE	RMSE	MAE
Random Forest	11.857950	3.4435374	2.1515895
PBRF [64]	9.8154816	3.1329669	2.1222036
RARTEN [4]	9.7350255	3.1201002	2.0906544
ECAPRAF-EN [5]	10.380147	3.2218236	2.1408014
ERDeR-MLP-EN (Ours)	**0.1505009**	**0.3879444**	**0.2750212**
ERDeR-WA-EN (Ours)	**0.1415230**	**0.3761954**	**0.2641758**
ERDeR-SA-EN (Ours)	**0.1205808**	**0.3472475**	**0.257047**

TABLE 18. Comparison of existing works with our proposed model on real estate valuation.

Algorithm	MSE	RMSE	MAE
Random Forest	108.0136379	10.39296098	5.251413432
PBRF [64]	105.5873882	10.2755724	5.250650213
RARTEN [4]	105.303779	10.26176296	5.281261517
ECAPRAF-EN [5]	30.33592275	5.507805621	4.207588771
ERDeR-MLP-EN (Ours)	**0.221604801**	**0.470749192**	**0.724839305**
ERDeR-WA-EN (Ours)	**0.227441072**	**0.476907824**	**0.36747779**
ERDeR-SA-EN (Ours)	**0.214130355**	**0.462742213**	**0.359331356**

ERDeR-SA-EN show that ERDeR-SA-EN with SA outperforms other two ensemble methods.

V. CONCLUSION AND FUTURE WORKS

In this study, a new hybrid model called ERDeR was presented, which contains three phases including base regression, shrinkage, and ensemble phases. In the base regression phase, the DRs were used as learners. In principle, in this paper, the learners were made of deep CNN whose number was equal to 30,40, and 50 parallel DRs. The next phase of the proposed model was shrinkage phase. In this phase, LASSO and EN methods were used to reduce the number of DRs and eliminate ineffective ones. It can be said that these methods operated as a pruner. Regarding the performance of two shrinkage methods in combination with ensemble methods, we can refer to the better performance of EN than LASSO. In addition, three non-shrinkage models were compared with new hybrid methods. The final phase of the model was the ensemble phase in which DRs were concatenated in the fusion phase and then aggregated using ensemble methods such as MLP, WA, and SA. Each method had different structures for the ensemble. In MLP, three fully-connected layers and the tanh activation function were used to non-linear transformation of the output. In the WA, one fully-connected layer and linear activation function were used for the linear transformation of the output. And in the SA method, no fully-connected layer was used and only SA was performed.

The obtained results showed that not only the type of shrinkage and ensemble methods were effective in decreasing the computational load in the test phase but also they achieved great performance. Among the shrinkage methods, EN provided better performance than LASSO. Concerning ensemble approaches, it can be referred to the better performance of MLP and WA. The WA approach due to having one fully-connected layer and performing a linear transformation on output, can be a more suitable method than other ensemble methods. In combining ensemble and shrinkage phases, WA-EN and MLP-EN models had better performance and high accuracy; while non-shrinkage methods did not have great performance. In the SA method, the performance of SA-NON was better than SA-EN and SA-LASSO. Therefore, SA which had a different structure cannot be a suitable method to improve the performance of the ERDeR model. The reason for this could be the structural weakness of SA in which simple averaging is used for the ensemble. Considering the measure of changes in different methods on simulation data, the reduction value of MLP and WA methods in combination with LASSO and EN, i.e. MLP-EN, MLP-LASSO, WA-EN, and WA-LASSO, was approximately equal to 18% and 14%, respectively. While the SA method was accompanied by almost 2% increase. This means that applying shrinkage methods and reducing DRs in MLP and WA approaches caused a decrease; while SA caused an increase. Regarding the obtained results from the three real datasets, like the simulation data, the WA and MLP approach in combination with the LASSO and EN methods provided better results than non-shrinkage models. By contrast, SA did not have a suitable performance. Overall, it can be concluded that the performance of shrinkage methods in combination with different ensemble approaches and different learners is better than non-shrinkage models.

As a future work, we will use deep forest algorithm instead of 1D-CNNs as a learner instead of DRs in this study and RF in [4] and [5]. In addition, different structure of ensemble methods might be used instead of MLP, WA, and SA.

REFERENCES

[1] F. Barboza, H. Kimura, and E. Altman, "Machine learning models and bankruptcy prediction," *Expert Syst. Appl.*, vol. 83, pp. 405–417, Oct. 2017, doi: 10.1016/j.eswa.2017.04.006.

[2] S. J. Thomas, M. L'Azou, A. D. T. Barrett, and N. A. C. Jackson, "Fast-track Zika vaccine development—Is it possible?" *New England J. Med.*, vol. 375, no. 13, pp. 1212–1216, Sep. 2016, doi: 10.1056/nejmp1609300.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[4] Z. Farhadi, H. Bevrani, and M.-R. Feizi-Derakhshi, "Improving random forest algorithm by selecting appropriate penalized method," *Commun. Statist.-Simul. Comput.*, vol. 2022, pp. 1–16, Nov. 2022, doi: 10.1080/03610918.2022.2150779.

[5] Z. Farhadi, H. Bevrani, M.-R. Feizi-Derakhshi, W. Kim, and M. F. Ijaz, "An ensemble framework to improve the accuracy of prediction using clustered random-forest and shrinkage methods," *Appl. Sci.*, vol. 12, no. 20, p. 10608, Oct. 2022, doi: 10.3390/app122010608.

[6] Z. Farhadi, H. Bevrani, and M.-R. Feizi-Derakhshi, "Combining regularization and dropout techniques for deep convolutional neural network," in *Proc. Global Energy Conf. (GEC)*, Oct. 2022, pp. 335–339, doi: 10.1109/GEC55014.2022.9986657.

[7] A. Foucquier, S. Robert, F. Suard, L. Stéphan, and A. Jay, "State of the art in building modelling and energy performances prediction: A review," *Renew. Sustain. Energy Rev.*, vol. 23, pp. 272–288, Jul. 2013, doi: 10.1016/j.rser.2013.03.004.

[8] R. Maclin, "Popular ensemble methods: An empirical study," *J. Artif. Intell. Res.*, vol. 11, pp. 169–198, Jul. 2016.

[9] D. Li, K. D. Ortegas, and M. White, "Exploring the computational effects of advanced deep neural networks on logical and activity learning for enhanced thinking skills," *Systems*, vol. 11, no. 7, p. 319, Jun. 2023, doi: 10.3390/systems11070319.

[10] Z. Wang, Y. Wang, and R. S. Srinivasan, "A novel ensemble learning approach to support building energy use prediction," *Energy Buildings*, vol. 159, pp. 109–122, Jan. 2018, doi: 10.1016/j.enbuild.2017.10.085.

[11] Z. Luo, H. Wang, and S. Li, "Prediction of international roughness index based on stacking fusion model," *Sustainability*, vol. 14, no. 12, p. 6949, Jun. 2022, doi: 10.3390/su14126949.

[12] M. Ramezani, M.-R. Feizi-Derakhshi, M.-A. Balafar, M. Asgari-Chenaghlu, A.-R. Feizi-Derakhshi, N. Nikzad-Khasmakhi, M. Ranjbar-Khadivi, Z. Jahanbakhsh-Nagadeh, E. Zafarani-Moattar, and T. Akan, "Automatic personality prediction: An enhanced method using ensemble modeling," *Neural Comput. Appl.*, vol. 34, no. 21, pp. 18369–18389, Nov. 2022, doi: 10.1007/s00521-022-07444-6.

[13] C. Fan, F. Xiao, and S. Wang, "Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques," *Appl. Energy*, vol. 127, pp. 1–10, Aug. 2014, doi: 10.1016/j.apenergy.2014.04.016.

[14] G. Giacinto and F. Roli, "Design of effective neural network ensembles for image classification purposes," *Image Vis. Comput.*, vol. 19, nos. 9–10, pp. 699–707, Aug. 2001, doi: 10.1016/s0262-8856(01)00045-2.

[15] A. C. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification," *Appl. Bioinf.*, vol. 2, no. 3, pp. 1–10, 2003.

[16] U. E. Laila, K. Mahboob, A. W. Khan, F. Khan, and W. Taekeun, "An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study," *Sensors*, vol. 22, no. 14, p. 5247, Jul. 2022, doi: 10.3390/s22145247.

[17] N. P. Tigga and S. Garg, "Prediction of type 2 diabetes using machine learning classification methods," *Proc. Comput. Sci.*, vol. 167, pp. 706–716, Jan. 2020, doi: 10.1016/j.procs.2020.03.336.

[18] D. J. Reddy, B. Mounika, S. Sindhu, T. P. Reddy, N. S. Reddy, G. J. Sri, K. Swaraja, K. Meenakshi, and P. Kora, "WITHDRAWN: Predictive machine learning model for early detection and analysis of diabetes," *Mater. Today, Proc.*, vol. 2020, pp. 1–15, Oct. 2020, doi: 10.1016/j.matpr.2020.09.522.

[19] D. M. Jose, A. M. Vincent, and G. S. Dwarakish, "Improving multiple model ensemble predictions of daily precipitation and temperature through machine learning techniques," *Sci. Rep.*, vol. 12, no. 1, pp. 1–25, Mar. 2022, doi: 10.1038/s41598-022-08786-w.

[20] M. Fallahian, E. Ahmadi, and F. Khoshnoudian, "A structural damage detection algorithm based on discrete wavelet transform and ensemble pattern recognition models," *J. Civil Struct. Health Monitor.*, vol. 12, no. 2, pp. 323–338, Apr. 2022, doi: 10.1007/s13349-021-00546-0.

[21] E. G. Moung, C. C. Wooi, M. M. Sufian, C. K. On, and J. A. Dargham, "Ensemble-based face expression recognition approach for image sentiment analysis," *Int. J. Electr. Comput. Eng.*, vol. 12, no. 3, p. 2588, Jun. 2022, doi: 10.11591/ijece.v12i3.pp2588-2600.

[22] J. Serey, M. Alfaro, G. Fuertes, M. Vargas, C. Durán, R. Ternero, R. Rivera, and J. Sabattin, "Pattern recognition and deep learning technologies, enablers of industry 4.0, and their role in engineering research," *Symmetry*, vol. 15, no. 2, p. 535, Feb. 2023, doi: 10.3390/sym15020535.

[23] M. Larsson, Y. Zhang, and F. Kahl, "Robust abdominal organ segmentation using regional convolutional neural networks," *Appl. Soft Comput.*, vol. 70, pp. 465–471, Sep. 2018, doi: 10.1016/j.asoc.2018.05.038.

[24] R. Pramanik, S. Dey, S. Malakar, S. Mirjalili, and R. Sarkar, "TOPSIS aided ensemble of CNN models for screening COVID-19 in chest X-ray images," *Sci. Rep.*, vol. 12, no. 1, pp. 1–19, Sep. 2022, doi: 10.1038/s41598-022-18463-7.

[25] A. Khatri, S. Agrawal, and J. M. Chatterjee, "Wheat seed classification: Utilizing ensemble machine learning approach," *Sci. Program.*, vol. 2022, pp. 1–9, Feb. 2022, doi: 10.1155/2022/2626868.

[26] J.-A. Lee and K.-C. Kwak, "Personal identification using an ensemble approach of 1D-LSTM and 2D-CNN with electrocardiogram signals," *Appl. Sci.*, vol. 12, no. 5, p. 2692, Mar. 2022, doi: 10.3390/app12052692.

[27] M. Asgari-Chenaghlu, M. R. Feizi-Derakhshi, L. Farzinvash, M. A. Balafar, and C. Motamed, "CWI: A multimodal deep learning approach for named entity recognition from social media using character, word and image features," *Neural Comput. Appl.*, vol. 34, no. 3, pp. 1905–1922, Feb. 2022, doi: 10.1007/s00521-021-06488-4.

[28] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.

[29] T. Adewumi, S. Sabah Sabry, N. Abid, F. Liwicki, and M. Liwicki, "T5 for hate speech, augmented data and ensemble," 2022, *arXiv:2210.05480*.

[30] K. Mnassri, P. Rajapaksha, R. Farahbakhsh, and N. Crespi, "BERT-based ensemble approaches for hate speech detection," in *Proc. IEEE Global Commun. Conf.*, Dec. 2022, pp. 4649–4654, doi: 10.1109/GLOBECOM48099.2022.10001325.

[31] Z. Jahanbakhsh-Nagadeh, M.-R. Feizi-Derakhshi, and A. Sharifi, "A deep content-based model for Persian rumor verification," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 21, no. 1, pp. 1–29, Jan. 2022, doi: 10.1145/3487289.

[32] D. R. Wijaya, F. Afianti, A. Arifianto, D. Rahmawati, and V. S. Kodogiannis, "Ensemble machine learning approach for electronic nose signal processing," *Sens. Bio-Sensing Res.*, vol. 36, Jun. 2022, Art. no. 100495, doi: 10.1016/j.sbsr.2022.100495.

[33] S. F. Stefenon, R. Bruns, A. Sartori, L. H. Meyer, R. G. Ovejero, and V. R. Q. Leithardt, "Analysis of the ultrasonic signal in polymeric contaminated insulators through ensemble learning methods," *IEEE Access*, vol. 10, pp. 33980–33991, 2022, doi: 10.1109/ACCESS.2022.3161506.

[34] S.-M. Chiu, Y.-C. Chen, and C. Lee, "Estate price prediction system based on temporal and spatial features and lightweight deep learning model," *Int. J. Speech Technol.*, vol. 52, no. 1, pp. 808–834, May 2021, doi: 10.1007/s10489-021-02472-6.

[35] T. Mahmud, M. A. Rahman, and S. A. Fattah, "CovXNet: A multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization," *Comput. Biol. Med.*, vol. 122, Jul. 2020, Art. no. 103869.

[36] M. M. Ahsan, T. E. Alam, T. Trafalis, and P. Huebner, "Deep MLP-CNN model using mixed-data to distinguish between COVID-19 and non-COVID-19 patients," *Symmetry*, vol. 12, no. 9, p. 1526, Sep. 2020, doi: 10.3390/sym12091526.

[37] Y. Liu, X. Yan, C.-A. Zhang, and W. Liu, "An ensemble convolutional neural networks for bearing fault diagnosis using multi-sensor data," *Sensors*, vol. 19, no. 23, p. 5300, Dec. 2019, doi: 10.3390/s19235300.

[38] L. Liu, N. Jia, L. Lin, and Z. He, "A cohesion-based heuristic feature selection for short-term traffic forecasting," *IEEE Access*, vol. 7, pp. 3383–3389, 2019, doi: 10.1109/ACCESS.2018.2889814.

[39] M. Cao, V. O. K. Li, and V. W. S. Chan, "A CNN-LSTM model for traffic speed prediction," in *Proc. IEEE 91st Veh. Technol. Conf. (VTC-Spring)*, May 2020, pp. 1–5, doi: 10.1109/VTC2020-Spring48590.2020.9129440.

[40] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[41] M. Canizo, I. Triguero, A. Conde, and E. Onieva, "Multi-head CNN–RNN for multi-time series anomaly detection: An industrial case study," *Neurocomputing*, vol. 363, pp. 246–260, Oct. 2019, doi: 10.1016/j.neucom.2019.07.034.

[42] M. E. Basiri, S. Nemati, M. Abdar, E. Cambria, and U. R. Acharya, "ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis," *Future Gener. Comput. Syst.*, vol. 115, pp. 279–294, Feb. 2021, doi: 10.1016/j.future.2020.08.005.

[43] J. Wang, L.-C. Yu, K. R. Lai, and X. Zhang, "Dimensional sentiment analysis using a regional CNN-LSTM model," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Stroudsburg, PA, USA, 2016, pp. 225–230, doi: 10.18653/v1/p16-2037.

[44] Z. M. Zain and N. M. Alturki, "COVID-19 pandemic forecasting using CNN-LSTM: A hybrid approach," *J. Control Sci. Eng.*, vol. 2021, pp. 1–23, Jul. 2021, doi: 10.1155/2021/8785636.

[45] X. Yu, H. Qiu, and S. Xiong, "A novel hybrid deep neural network to predict pre-impact fall for older people based on wearable inertial sensors," *Frontiers Bioeng. Biotechnol.*, vol. 8, pp. 1–10, Feb. 2020, doi: 10.3389/fbioe.2020.00063.

[46] P. Thomas, H. B. E. Haouzi, M.-C. Suhner, A. Thomas, E. Zimmermann, and M. Noyel, "Using a classifier ensemble for proactive quality monitoring and control: The impact of the choice of classifiers types, selection criterion, and fusion process," *Comput. Ind.*, vol. 99, pp. 193–204, Aug. 2018, doi: 10.1016/j.compind.2018.03.038.

[47] S. Ma and F. Chu, "Ensemble deep learning-based fault diagnosis of rotor bearing systems," *Comput. Ind.*, vol. 105, pp. 143–152, Feb. 2019, doi: 10.1016/j.compind.2018.12.012.

[48] B.-K. Kim, H. Lee, J. Roh, and S.-Y. Lee, "Hierarchical committee of deep CNNs with exponentially-weighted decision fusion for static facial expression recognition," in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 427–434, doi: 10.1145/2818346.2830590.

[49] A. Baldominos, Y. Saez, and P. Isasi, "Model selection in committees of evolved convolutional neural networks using genetic algorithms," in *Intelligent Data Engineering and Automated Learning—IDEAL 2018* (Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Cham, Switzerland: Springer, 2018, pp. 364–373, doi: 10.1007/978-3-030-03493-1_39.

[50] M. S. Haghighi, A. Vahedian, and H. S. Yazdi, "Creating and measuring diversity in multiple classifier systems using support vector data description," *Appl. Soft Comput.*, vol. 11, no. 8, pp. 4931–4942, Dec. 2011, doi: 10.1016/j.asoc.2011.06.006.

[51] Z.-H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: Many could be better than all," *Artif. Intell.*, vol. 137, nos. 1–2, pp. 239–263, May 2002, doi: 10.1016/s0004-3702(02)00190-x.

[52] Y.-W. Kim and I.-S. Oh, "Classifier ensemble selection using hybrid genetic algorithms," *Pattern Recognit. Lett.*, vol. 29, no. 6, pp. 796–802, Apr. 2008, doi: 10.1016/j.patrec.2007.12.013.

[53] K. Wang, Y. An, J. Zhou, Y. Long, and X. Chen, "A novel multi-level feature selection method for radiomics," *Alexandria Eng. J.*, vol. 66, pp. 993–999, Mar. 2023, doi: 10.1016/j.aej.2022.10.069.

[54] A. Zaki, A. Métwalli, M. H. Aly, and W. K. Badawi, "Enhanced feature selection method based on regularization and kernel trick for 5G applications and beyond," *Alexandria Eng. J.*, vol. 61, no. 12, pp. 11589–11600, Dec. 2022, doi: 10.1016/j.aej.2022.05.024.

[55] X. Zhou, X. Liu, G. Zhang, L. Jia, X. Wang, and Z. Zhao, "An iterative threshold algorithm of log-sum regularization for sparse problem," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4728–4740, Sep. 2023, doi: 10.1109/TCSVT.2023.3247944.

[56] H. Wang, Y. Shen, S. Wang, T. Xiao, L. Deng, X. Wang, and X. Zhao, "Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease," *Neurocomputing*, vol. 333, pp. 145–156, Mar. 2019, doi: 10.1016/j.neucom.2018.12.018.

[57] R. Daw and C. K. Wikle, "REDS: Random ensemble deep spatial prediction," *Environmetrics*, vol. 34, no. 1, Feb. 2023, Art. no. e2780, doi: 10.1002/env.2780.

[58] S. Wang, Y. Wu, R. Li, and X. Wang, "Remote sensing-based retrieval of soil moisture content using stacking ensemble learning models," *Land Degradation Develop.*, vol. 34, no. 3, pp. 911–925, Feb. 2023, doi: 10.1002/ldr.4505.

[59] A. Akbas and S. Buyrukoglu, "Stacking ensemble learning-based wireless sensor network deployment parameter estimation," *Arabian J. Sci. Eng.*, vol. 48, no. 8, pp. 9739–9748, Aug. 2023, doi: 10.1007/s13369-022-07365-5.

[60] Y. J. Cruz, M. Rivas, R. Quiza, A. Villalonga, R. E. Haber, and G. Beruvides, "Ensemble of convolutional neural networks based on an evolutionary algorithm applied to an industrial welding process," *Comput. Ind.*, vol. 133, Dec. 2021, Art. no. 103530, doi: 10.1016/j.compind.2021.103530.

[61] A. Arjmand, O. Tsakai, V. Christou, A. T. Tzallas, M. G. Tsipouras, R. Forlano, P. Manousou, R. D. Goldin, C. Gogos, E. Glavas, and N. Giannakeas, "Ensemble convolutional neural network classification for pancreatic steatosis assessment in biopsy images," *Information*, vol. 13, no. 4, p. 160, Mar. 2022, doi: 10.3390/info13040160.

[62] A. Yazdizadeh, Z. Patterson, and B. Farooq, "Ensemble convolutional neural networks for mode inference in smartphone travel survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 6, pp. 2232–2239, Jun. 2020, doi: 10.1109/TITS.2019.2918923.

[63] M. Shi, W. Hu, M. Li, J. Zhang, X. Song, and W. Sun, "Ensemble regression based on polynomial regression-based decision tree and its application in the in-situ data of tunnel boring machine," *Mech. Syst. Signal Process.*, vol. 188, Apr. 2023, Art. no. 110022, doi: 10.1016/j.ymssp.2022.110022.

[64] H. Wang and G. Wang, "Improving random forest algorithm by lasso method," *J. Stat. Comput. Simul.*, vol. 91, no. 2, pp. 353–367, Jan. 2021, doi: 10.1080/00949655.2020.1814776.

[65] K. Faber, M. Pietron, and D. Zurek, "Ensemble neuroevolution-based approach for multivariate time series anomaly detection," *Entropy*, vol. 23, no. 11, p. 1466, Nov. 2021, doi: 10.3390/e23111466.

[66] A. Manna, R. Kundu, D. Kaplun, A. Sinitca, and R. Sarkar, "A fuzzy rank-based ensemble of CNN models for classification of cervical cytology," *Sci. Rep.*, vol. 11, no. 1, pp. 1–18, Jul. 2021, doi: 10.1038/s41598-021-93783-8.

[67] D. Xue, X. Zhou, C. Li, Y. Yao, M. M. Rahaman, J. Zhang, H. Chen, J. Zhang, S. Qi, and H. Sun, "An application of transfer learning and ensemble learning techniques for cervical histopathology image classification," *IEEE Access*, vol. 8, pp. 104603–104618, 2020, doi: 10.1109/ACCESS.2020.2999816.

[68] X. Lu, C. Ma, B. Ni, X. Yang, I. Reid, and M.-H. Yang, "Deep regression tracking with shrinkage loss," *Eur. Conf. Comput. Vis.*, vol. 11218, pp. 369–386, 2018, doi: 10.1007/978-3-030-01264-9.

[69] V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab, "Robust optimization for deep regression," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2830–2838, doi: 10.1109/ICCV.2015.324.

[70] S. Lathuilière, P. Mesejo, X. Alameda-Pineda, and R. Horaud, "A comprehensive analysis of deep regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2065–2081, Sep. 2020, doi: 10.1109/TPAMI.2019.2910523.

[71] S.-P. Mousavi, R. Nakhaei-Kohani, S. Atashrouz, F. Hadavimoghaddam, A. Abedi, A. Hemmati-Sarapardeh, and A. Mohaddespour, "Modeling of H2S solubility in ionic liquids: Comparison of white-box machine learning, deep learning and ensemble learning approaches," *Sci. Rep.*, vol. 13, no. 1, pp. 1–23, May 2023, doi: 10.1038/s41598-023-34193-w.

[72] M. Bhuiyan and M. S. Islam, "A new ensemble learning approach to detect malaria from microscopic red blood cell images," *Sensors Int.*, vol. 4, 2023, Art. no. 100209, doi: 10.1016/j.sintl.2022.100209.

[73] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: 10.1038/323533a0.

[74] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural Netw.*, vol. 12, no. 1, pp. 145–151, Jan. 1999, doi: 10.1016/s0893-6080(98)00116-6.

[75] J. C. Duchi, P. L. Bartlett, and M. J. Wainwright, "Randomized smoothing for (parallel) stochastic optimization," in *Proc. IEEE 51st IEEE Conf. Decis. Control (CDC)*, Dec. 2012, pp. 5442–5444, doi: 10.1109/CDC.2012.6426698.

[76] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[77] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, vol. 17, no. 2, p. 425, Feb. 2017, doi: 10.3390/s17020425.

[78] S. Huang, J. Tang, J. Dai, and Y. Wang, "Signal status recognition based on 1DCNN and its feature extraction mechanism analysis," *Sensors*, vol. 19, no. 9, p. 2018, Apr. 2019, doi: 10.3390/s19092018.

[79] T. Zan, H. Wang, M. Wang, Z. Liu, and X. Gao, "Application of multi-dimension input convolutional neural network in fault diagnosis of rolling bearings," *Appl. Sci.*, vol. 9, no. 13, p. 2690, Jul. 2019, doi: 10.3390/app9132690.

[80] Y. Zhang, L. Zhang, and M. A. Hossain, "Adaptive 3D facial action intensity estimation and emotion recognition," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1446–1464, Feb. 2015, doi: 10.1016/j.eswa.2014.08.042.

[81] S. C. Neoh, L. Zhang, K. Mistry, M. A. Hossain, C. P. Lim, N. Aslam, and P. Kinghorn, "Intelligent facial emotion recognition using a layered encoding cascade optimization model," *Appl. Soft Comput.*, vol. 34, pp. 72–93, Sep. 2015, doi: 10.1016/j.asoc.2015.05.006.

[82] D. W. Opitz and J. W. Shavlik, "Actively searching for an effective neural network ensemble," *Connection Sci.*, vol. 8, nos. 3–4, pp. 337–354, Dec. 1996, doi: 10.1080/095400996116802.

[83] M. P. Perrone and L. N. Cooper, "When networks disagree: Ensemble methods for hybrid neural networks," in *World Scientific Series in 20th Century Physics*, vol. 10, 1995, pp. 342–358, doi: 10.1142/9789812795885_0025.

[84] L. Wang, H. Liu, Z. Pan, D. Fan, C. Zhou, and Z. Wang, "Long short-term memory neural network with transfer learning and ensemble learning for remaining useful life prediction," *Sensors*, vol. 22, no. 15, p. 5744, Aug. 2022, doi: 10.3390/s22155744.

[85] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Society: Ser. B (Methodological)*, vol. 58, no. 1, pp. 267–288, Jan. 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[86] W. Dai, X. Zhou, D. Li, S. Zhu, and X. Wang, "Hybrid parallel stochastic configuration networks for industrial data analytics," *IEEE Trans. Ind. Informat.*, vol. 18, no. 4, pp. 2331–2341, Apr. 2022, doi: 10.1109/TII.2021.3096840.

[87] X. Liu, S. Wang, S. Lu, Z. Yin, X. Li, L. Yin, J. Tian, and W. Zheng, "Adapting feature selection algorithms for the classification of Chinese texts," *Systems*, vol. 11, no. 9, p. 483, Sep. 2023, doi: 10.3390/systems11090483.

[88] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Stat. Soc. Ser. B: Stat. Methodology*, vol. 67, no. 2, pp. 301–320, Apr. 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[89] D. Harrison and D. L. Rubinfeld, "Hedonic housing prices and the demand for clean air," *J. Environ. Econ. Manage.*, vol. 5, no. 1, pp. 81–102, Mar. 1978, doi: 10.1016/0095-0696(78)90006-2.

[90] C.-S.-J. Chu, "Time series segmentation: A sliding window approach," *Inf. Sci.*, vol. 85, nos. 1–3, pp. 147–173, Jul. 1995, doi: 10.1016/0020-0255(95)00021-g.

ZARI FARHADI received the B.Sc. and M.Sc. degrees in mathematical statistics from the University of Tabriz, Tabriz, Iran, where she is currently pursuing the Ph.D. degree. Her research interests include high dimensional data, machine learning, deep learning, and data science.

MOHAMMAD-REZA FEIZI-DERAKHSHI received the B.Sc. degree in software engineering from the University of Isfahan, Isfahan, Iran, and the M.Sc. and Ph.D. degrees in artificial intelligence from Iran University of Science and Technology. He is currently a Full Professor of computer engineering with the University of Tabriz. His research interests include natural language processing, optimization algorithms, social network analysis, and intelligent databases.

HOSSEIN BEVRANI received the Ph.D. degree in mathematical statistics from Moscow State University, Moscow, Russia. He is currently a Full Professor of statistics with the University of Tabriz. He has been involved in several national and international projects with many partners, in the position of main team member or leader. He has published peer-reviewed articles in both theoretical studies and real applications. His research interests include statistical multivariate methods, generalized linear method, simulation, high dimensional data, and model selection.

WONJOON KIM (Member, IEEE) received the Ph.D. degree in industrial engineering from Seoul National University, South Korea. He is currently an Assistant Professor with Dongduk Women's University, specializing in data science, machine learning methods, and their applications to industrial, business, and financial problems. His research interests include machine learning methods and applications to industry problems.

MUHAMMAD FAZAL IJAZ was a Visiting Guest Professor and an Assistant Professor with tertiary institutes, including Dongguk University; Technology de Monterrey, Campus Mexico City and Guadalajara, Mexico; Sejong University, Seoul, and The University of Melbourne, Australia. From 2021 to 2023, each year he was presented among "Top 2% Scientists in the World" by Stanford University for his career achievements. He has published numerous research articles in several international peer-reviewed journals, including IEEE Transactions on Industrial Informatics, IEEE Internet of Things Journal, *Scientific Reports*, *Cancers*, *Human-Centric Computing and Information Sciences*, IEEE, *Biomedical Signal Processing and Control*, and *Computational Intelligence*. His research interests include human-centered AI, medical image analysis, medical artificial intelligence, the Internet of Things, and data mining. He is a reviewer and an editorial board member of numerous top ranked journals.