Scientia Iranica (2024) 31(5), 417–429

SCIENTIA IRANICA

Sharif University of Technology
Scientia Iranica
Transactions D: Computer Science & Engineering and Electrical Engineering
https://scientiairanica.sharif.edu

Optimized deep networks structure to improve the accuracy of estimator algorithm in deep networks learning

H. Rezaei Nezhad[a], F. Keynia[b,*], and A. Sabbagh Molahosseini[a,1]

a. *Department of Computer, Kerman Branch, Islamic Azad University, Kerman, Iran.*
b. *Department of Energy Management and Optimization, Institute of Science and High Technology and Environmental Sciences, Graduate University of Advanced Technology, Kerman, Iran.*
1. *Present address: School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, UK.*

Received 20 May 2023; received in revised form 9 August 2023; accepted 22 August 2023

KEYWORDS
Optimization algorithm;
Time series;
Estimation;
Prediction;
Convolutional neural network;
Long short-term memory.

Abstract. The present study considers decreasing prediction error for various types of time series and the uncertainty in estimation parameters, improving the structure of the Deep Neural Network (DNN), and increasing response speed in the proposed neural network method. Additionally, the competitive performance and the collaboration among the neurons of the DNN are also increased. The selected data is related to weather prediction for Qeshm, which has suitable weather conditions for our study, spanning from 2016 onwards. In this study, to analyze the prediction issue of power consumption of domestic expenses in the indefinite and severe fluctuation mode, we decided to combine the two methods of Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN). For the training of the deep network, the Back Propagation (BP) algorithm is used. The results indicate that Gated Recurrent Unit (GRU) networks compared to other models (Multi-Layer Perceptrons (MLP), CNN, and DNN) produce more realistic results, and also two-way networks obtained better results on test data compared to LSTM networks. Root Mean Square Error (RMSE) prediction is more realistic than the LSTM model on test and training data against significant data. A GRU network has two gates of rt readjust and Z_t forgetting, which helps to ensure that long-term dependencies of gradient fading will not occur.

1. Introduction

At its simplest, a neural network with some level of complexity, usually at least two layers, qualifies as a Deep Neural Network (DNN), or deep net for short. Deep nets process data in complex ways by employing sophisticated math modeling. The optimization algorithm is formed based on training and learning and the process of training and learning in a class that was introduced for the first time by Rao et al. [1].

The uniformly minimum variance unbiased es-

*. *Corresponding author.*
E-mail addresses: rezaeii.Iauqeshm@gmail.com (H. Rezaei Nezhad); f.keynia@kgut.ac.ir (F. Keynia); Sabbagh@iauk.ac.ir and a.sabbaghmolahosseini@qub.ac.uk (A. Sabbagh Molahosseini)

To cite this article:
H. Rezaei Nezhad, F. Keynia, and A. Sabbagh Molahosseini "Optimized deep networks structure to improve the accuracy of estimator algorithm in deep networks learning", *Scientia Iranica* (2024) **31**(5), pp. 417-429
DOI: 10.24200/sci.2023.62337.7782

timator is one of the most fundamental and important estimation methods in classical statistics, but its existence and characterization are usually challenging to investigate when one moves beyond exponential families. In the past several decades, many shrinkage estimations, regression, and variable selection methods have been proposed. For instance, the James-Stein estimator dominates the maximum likelihood estimator in terms of expected total squared loss beyond two-dimensional Gaussian models [2].

Deep Learning (DL) is part of a broader family of Machine Learning (ML) methods, which is based on Artificial Neural Networks (ANN) with representation learning. A new efficient optimization method, called 'Teaching-Learning-Based Optimization (TLBO)', is for the optimization of mechanical design problems. This method works on the effect of the influence of a teacher on learners. The TLBO algorithm utilizes the ability of students in the classroom and the teaching of class teachers to improve the educational level of the class. Teachers and students are the two main elements of the TLBO algorithm. TLBO is a powerful meta-heuristic algorithm for solving complex optimization problems about the global optimum. Many TLBO variants have been presented to improve the local optima avoidance capability and to increase the convergence speed. Due to the rapid development of engineering requirements, global optimization has attracted a lot of interest. In general, there are two classes of methods for solving global optimization problems, i.e., nature-inspired methods and deterministic methods. Because the deterministic method often fails in solving complex global optimization problems, nature-inspired methods are more popular. Meta-heuristic Algorithms (MAs), which are instinctively immune from non-smooth behaviors, are the most widely used nature-inspired methods. In the past ten years, many state-of-the-art MAs, such as Aquila Optimizer (AO), Reptile Search Algorithm (RSA), Moth Flame Optimization (MFO), Arithmetic Optimization Algorithm (AOA), Whale Optimization Algorithm (WOA), TLBO, Sine Cosine Algorithm (SCA), etc., have been developed. Because of its simple structure, lack of specific parameters, and strong practicability, TLBO has been widely used in many engineering applications, such as personalized recommender systems, controller design, wind power forecasting, parameter estimation, neural network training, and so on. Accordingly, the teacher phase and student phase create two important and fundamental parts of this algorithm. The algorithm output is the scores of students and their level of knowledge in which the quality and ability of teachers in this field are consequential. The TLBO algorithm is a modern optimization algorithm based on a population, where this population consists of the members of the class.

The DNN is a type of feedforward neural network, and the connection pattern among its neurons is inspired by the visual cortex of animals' brains. The architecture of this network is based on the following concepts:

1. Separated connections;
2. Common weights;
3. The deletion of some neurons using the layers with the name of the merged layer.

Neural networks have significant features which are considered as follows:

- **Learning capability:** Learning is the capability to regulate network parameters (synaptic weights) to receive information to be reminded of in the future. If the network is trained for a specific status, and a small change in the network environmental situation (special condition) occurs, the network will be efficient for the new situation with brief training. There are weight changes, which are called learning laws.
- **Generalization capability:** Learning, in general, manages a comprehensive range of the set of inputs and outputs. The network has learning capability and generalization power to the whole set, and it can perform correctly for patterns of undefined input-output.
- **Parallel processing:** The general duty of processing is distributed among smaller and independent processors, which causes an increase in the speed of processing.

Studying variables in the research explained as follows:

- **Dependent variable:** The weights of DNN layers.
- **Independent variable:** The data of time series for training the data network The present study considers finding a suitable answer to the following questions:
  1. Which part of the Deep Network (DN) structure can be improved by optimization algorithms?
  2. Does the estimator based on smart methods such as learning have any better performance than classic estimators?
  3. Do Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) methods cause the decrease of the prediction error of power consumption of household charges in indefinite and severe fluctuation mode?

In this regard, the following hypotheses are considered:

- All-weather features such as the least humidity, the most humidity, humidity average, etc. will be available;

- There will not be any dependence among the dimensions;
- The CNN has three inputs.

An optimization algorithm based on training and learning is formed based on the process of training and learning in a class. A DNN is a type of feedforward neural network whose connection pattern among its neurons is inspired by the visual cortex of animals' brains. The present study considers decreasing prediction error for the types of time series and the uncertainty in estimation parameters, improving the structure of the DNN, and increasing response speed in the proposed neural network method; additionally, the competitive performance and the collaboration among the neurons of the DNN are also increased.

2. Literature review

2.1. Analysis algorithm of main components

The Principal Component Analysis (PCA) technique is the best method for decreasing data dimensions in linear mode. Less important coefficients obtained from this transform are deleted, and lost data is less than other methods. PCA-CNN (Principal Component Analysis-Convolutional Neural Network) is a development method of the CNN method which gives special treatment to the dimension reduction process in the input data. The dimension reduction process is carried out by utilizing the PCA method so that the data processing process becomes faster without losing important information, obtaining better method performance. The PCA-CNN method is implemented on a dataset from the Situbondo district, which is classified into five land cover classes. Of course, PCA applications are not limited to the decrease of dimensions; they are also used in other fields such as pattern recognition and facial recognition. In this method, new coordinates are defined for the data, and the data is noted based on these new coordinates. The first axis should be in the direction where data variance is maximum (where data separation is highest). Gorgoglione et al. [3] adopted PCA to assess the effect of rainfall, watershed, and drainage network characteristics on urban nutrient runoff in poorly gauged areas. The second axis should be perpendicular to the first axis so that data variance is maximum. Hence, the next axes, perpendicular to all previous axes, are such that the data have the most separation [4].

2.1.1. Accuracy of estimator algorithm in DNs learning

An estimation algorithm is an algorithm that produces an estimate for some value (such as the length of a stream). Similarly, an approximation algorithm is an algorithm that given an optimization problem finds a solution that approximately optimizes the objective function of the problem. Direction Of Arrival (DOA) estimation, also named spatial spectrum estimation, estimates the direction angle of the spatial signal reaching the array reference element by processing the received signal of the array. Traditional DOA estimation methods are mainly based on beamforming [5].

CNN-based DL is a powerful, recently developed image classification approach. With origins in the computer vision and image processing communities, the accuracy assessment methods developed for CNN-based DL use a wide range of metrics that may be unfamiliar to the Remote Sensing (RS) community. Array signal processing, also known as spatial domain signal processing, is an important branch of the signal processing field, widely used in radar signals, underwater sonar, wireless communication, radio astronomy, and other fields [6]. Mainly, to process the signals received by the array, the useful signals must be enhanced, useless interference and noise must be suppressed, and important parameters must be obtained. Estimation of the DOA is one of the important research contents of array signal processing. DOA estimation, also named spatial spectrum estimation, estimates the direction angle of the spatial signal reaching the array reference element by processing the received signal of the array [7].

Estimation of Distribution Algorithms (EDAs) are a novel class of evolutionary optimization algorithms that were developed as a natural alternative to Genetic Algorithms (GAs) in the last decade. The principal advantages of EDAs over GAs are the absence of multiple parameters to be tuned (e.g., crossover and mutation probabilities) and the expressiveness and transparency of the probabilistic model that guides the search process. In addition, EDAs have been proven to be better suited to some applications than GAs, while achieving competitive and robust results in the majority of tackled problems. EDAs belong to the class of evolutionary algorithms. The main difference between EDAs and most conventional evolutionary algorithms is that evolutionary algorithms generate new candidate solutions using an implicit distribution defined by one or more variation operators, whereas EDAs use an explicit probability distribution encoded by a Bayesian network, a multivariate normal distribution, or another model class. Similar to other evolutionary algorithms, EDAs can be used to solve optimization problems defined over a number of representations from vectors to LISP style S expressions, and the quality of candidate solutions is often evaluated using one or more objective functions [8].

The CNN is a common DL algorithm, which is a feedforward neural network that can reduce the number of parameters to a large extent by local connectivity and weight sharing. A typical CNN model usually consists of several convolutional and pooling layers connected alternately, ending with a fully connected layer [9]. DL applications are used in industries from automated driving to medical devices. Automated driving: Automotive researchers are using DL to automatically detect objects such as stop signs and traffic lights. In addition, DL is used to detect pedestrians, which helps decrease accidents. LeNet is the first successful application of CNNs and was developed by Yann Lecun in the 1990s that was used to read zip codes, digits, etc. The latest work is called LeNet-5, which is a 5-layer CNN that reaches 99.2% accuracy on isolated character recognition [10].

2.2. Convolution Neural Network (CNN)

The neural network appears as a practical technology that is successfully utilized in different fields such as classification, estimation, speech recognition, drug recognition, image processing, and signal processing. There are three types of layers that make up the CNN: Convolutional layers, pooling layers, and Fully-Connected (FC) layers. When these layers are stacked, a CNN architecture is formed. The most important advantage of neural networks is their self-adaptive and self-organizing capability, and their proactive operation, etc. In created neural network models, progressive multilayer networks are generally used with the Back Propagation (BP) training algorithm for classification, which is one of the supervised training methods. The structure of this network includes in an input layer, a middle layer, and an output layer. There is one or more processor element (neuron) in each layer, which is related to all neurons of the next layer with weighted connections. The vector of input data of the model is mapped to the neurons of the first layer. In this layer, there is no process, and the neurons of the output layer are mapped to the output vector of the model. The number of neurons in the input and output layers depends on the number of input and output variables of the model. Still, the selection of the number of neurons in the middle layer is determined by trial and error.

2.3. Advantages of CNNs

One of the main advantages of CNNs is that they can learn from raw pixel data, without requiring any manual feature engineering or preprocessing. This means that they can automatically discover and adapt to the most salient characteristics of images, such as edges, shapes, colors, textures, and objects. This also reduces the dimensionality and complexity of the input data, making the training and inference faster and more efficient. Another advantage of CNNs is that they can exploit the spatial and hierarchical structure of the images by using filters that preserve the local connectivity and context of the pixels, and by building more abstract and high-level representations as they go deeper into the network. This allows them to capture the variability and diversity of the images, and to generalize well to new and unseen data.

Wood investigated a ML approach for three objectives, one of which is predicting the location of targets. This work performed two-dimensional positioning with a circular object using the K-Nearest Neighbors (KNN) algorithm and a homogeneous non-magnetic wall. Common methods used in Through-the-Wall Radar (TWR) for locating targets, and estimating wall parameters, such as permittivity, wall thickness, etc., are performed separately [11].

DNs consist of two steps: The generation step and the pruning step. In the generation step, the network generates hidden layers layer by layer until accuracy reaches the threshold. The network uses a pruning algorithm based on Hebb's rule or Pearson's correlation for adaptation in the pruning step. Experimental results show that compared with traditional neural network topology optimization algorithms, Genetic Algorithm to optimize DNs (GA-DNAs) can generate neural networks with higher construction efficiency, lower structure complexity, and higher classification accuracy. Structural optimization has qualified as an important tool in the design process over the past decades. This method can be grouped into topology, size, and shape optimization. The objective of this optimization can be to minimize the stress weight or compliance for a given amount of material and boundary conditions. A wide range of algorithms is used to build the optimal neural network structure. The first of these algorithms is the tiled constructing algorithm. The idea of this algorithm is to add new layers of neurons in a way that input training vectors that have different respective initial values, would have a different internal representation in the algorithm. Another prominent representative is the fast superstructure algorithm [12].

According to this algorithm, new neurons are added between the output layers. The role of these neurons is to correct the output neurons' errors. In general, a neural network that is based on this algorithm has the form of a binary tree.

2.4. Research literature

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature [13].

The preliminary theoretical base for contemporary neural networks was independently proposed by Bain [14] and Brush [15]. In their work, both thoughts and body activity resulted from interactions among neurons within the brain.

DNNs can recognize voice commands, identify voices, recognize sounds and graphics, and do much more than a neural network. Deep learning networks utilize "big data" along with algorithms to solve a problem, and these DNNs can solve problems with limited or no human input.

For Bain, every activity led to the firing of a certain set of neurons. When activities were repeated, the connections between those neurons strengthened. According to his theory, this repetition was what led to the formation of memory. The general scientific community at the time was skeptical of Bain's theory because it required what appeared to be an inordinate number of neural connections within the brain. It is now apparent that the brain is exceedingly complex and that the same brain "wiring" can handle multiple problems and inputs.

James's theory was similar to Bain's; however, he suggested that memories and actions resulted from electrical currents flowing among the neurons in the brain. His model, by focusing on the flow of electrical currents, did not require individual neural connections for each memory or action.

Sherrington conducted experiments to test James's theory. He ran electrical currents down the spinal cords of rats. However, instead of demonstrating an increase in electrical current as projected by James, Sherrington found that the electrical current strength decreased as the testing continued over time. Importantly, this work led to the discovery of the concept of habituation [16].

Li et al., after surveying the reasons of overfitting and unsuitable generalization of neural networks, by applying changes in the neural network and using a class of Radial Basis Function (RBF) structuring delay of neural networks, succeeded in creating a neural network with high accuracy and with a lower number of neurons in the hidden layer of the neural network and tested its results in the actual world [17].

Lahmiri, by applying Discrete Wavelet Transform (DWT) and dividing the price time series into major and minor parts, concluded that the major part has lower frequency and separation, and it is suitable for long-term prediction. After applying the noted transformations and using feedback neural networks, he considered stock price prediction, and by investigating his theory in 15 databases, he concluded that his proposed algorithm has better performance than RW and Autoregressive Moving-Average (ARMA) models [18].

Ticknor, considering daily prices and technical analysis indicators as the inputs of the neural network, predicts the total price of the next day. Noting the presented complexities in the trend of stock price changes and the problems for its prediction, to prevent overfitting and overtraining, Ticknor suggests that the neural network is controlled by the Bayesian algorithm. For models with high complexity, fines should be determined to prevent overfitting and overtraining. He tested the accuracy of his claim on Microsoft and Goldman stocks [19].

Kara et al., noting that the prediction of stock price changes is a considerable problem and it being beneficial if accurate, declare that the expansion of mathematical models of this magnitude to address the intrinsic complexities of the stock market is difficult. Using technical analysis indicators as input, they investigate the performance of two algorithms: neural network classification and supports vector machines. By comparing the results, they found the performance of neural networks better and more suitable [20].

Zhang et al., developed an Support Vector Machine (SVM)-based method for two-dimensional locating under a homogeneous wall and a circular metal cylinder object. They also attempted to estimate the wall parameters using the same method, which is based on SVM [21].

In the article of Kose and Arslan, a method is proposed for training ANFIS neural networks, which have the capability of updating easier and faster than the Gradient method. In this method, which is called chaotic particle swarm intelligence, there is no need for learning rate, and the proposed learning algorithms are a combination of the descending Gradient method and the squared difference method with swarm intelligence algorithms such as the GA and the Ant colony algorithm, etc. [22].

In the article by Ma et al., a combined neural-fuzzy method is presented based on monthly different charge patterns for predicting the consumption charges of distribution companies. In the article, the mathematical equations related to the calculations, hypotheses, the type of neural network, the applied fuzzy logic, and the proposed algorithm are presented. This expert system is based on different statuses and times by considering decided input data using each one of the patterns [23].

A fully connected deep LSTM network is proposed by Zhu et al. [24] to recognize action with a framework composed of three LSTM and two feed-forward layers, incorporating co-occurrence regularization into the loss function, to explore the conjunctions of discriminative joints and different co-occurrences for several actions. Li et al. [17] propose a deep LSTM framework based on Recurrent Neural Network (RNN) to better localize the start and end of actions with a regression module to automatically extract features. This joint classification-regression RNN considers the sequence frame by frame and does not require a sliding window approach. Zhang et al. [25] presents a hierarchical approach, proposing three exploration fusion methods based on multilayer LSTM.

Some authors use a combination of CNN and LSTM to extract spatiotemporal information. However, unlike our approach, they merge the individual scores obtained from the CNN and the LSTM. Also, contrary to the method proposed by us, they consider all the joints of the skeleton, extrapolating other information such as distance and trajectory between the joints and the poses.

Effective utilization of present data in neural network trainer signals is among the most efficient cases on network performance. In the proposed method by Qiu et al., a suitable selection of training signals is made through their separation using the experimental mode method. The results of handled simulations indicate that using the proposed method leads to positive and considerable results [26].

3. Research method

The considered dataset is related to weather prediction for Qeshm, which has suitable weather conditions for our study, spanning from 2016 onwards. This dataset has nine features. In this dataset, features such as year, solar day, solar month, dry temperature average, etc. are used for the predicting the average humidity. The objective of the research is to estimate the pollution rate of CO_2. In this research, using an analysis algorithm of major data components, the present space is conveyed to a new space, and the dependence among the dimensions is lost. The accuracy of the estimation process is increased. In this study, a CNN is used for estimating the pollution rate of CO_2. To improve the accuracy of the CNN, an optimization algorithm based on training and learning in the training process is used to determine the weights and biases of the CNN in an optimized manner. Dimensional reduction is a process carried out to simplify existing variables, reducing their number without losing the information contained in the initial data. One of the methods used in dimension reduction is PCA. The function of PCA is to change the initial variables, reducing as many as n variables to k new variables called Principal Components (PCs). Assuming that the CNN includes three inputs in the input layer, four neurons in the hidden layer, and one neuron in the output layer, the process of neural network training using an optimization algorithm based on training and learning includes two major stages: The primary preparation stage and the repetition stage, which are explained in the following parts, along with the details of neural network training (Figure 1).

The primary preparation stage in the optimization algorithm is based on training and learning.

In this stage, a population of students is made, where each student comprises two components: Decision-making (status) variables and an objective function. Decision variables are considered for each student of neural network weights, and the Mean Squared Error (MSE) is considered for the training data as the objective function. Figure 2 shows the status of a student, which represents the weights of the CNN.

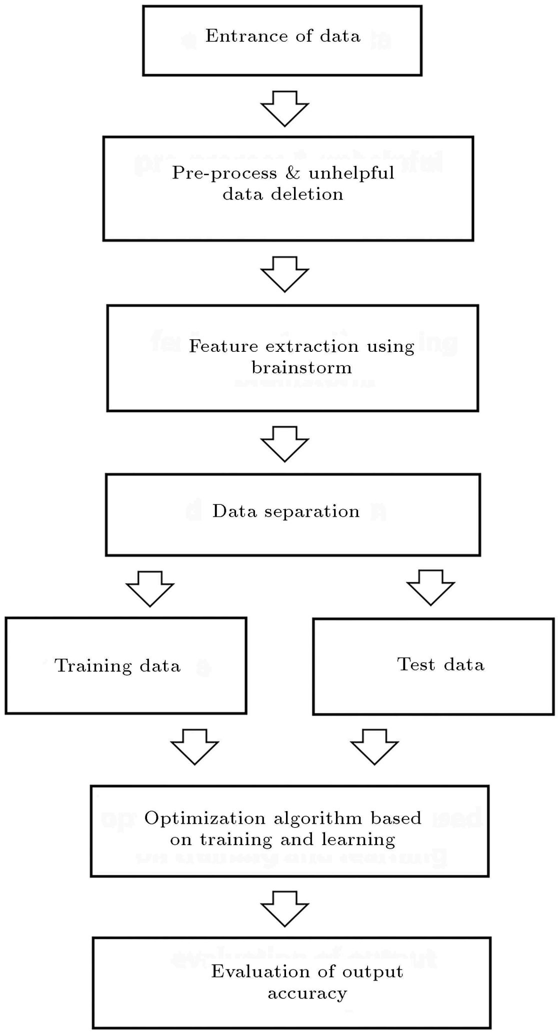

Figure 1. Proposed algorithm flowchart for neural network training.

In this stage, the values of the status variables for all students are randomized, and the MSE of the neural network is calculated for each member of the population.

The repetition stage in the optimization algorithm is based on training and learning.

In this stage, the following operations are repeated to fulfill the finishing situation:

- **Mean calculation:** In this step, the status of students is averaged, including the values of weights of the CNN;
- **Recognition of teacher:** In this step, the best

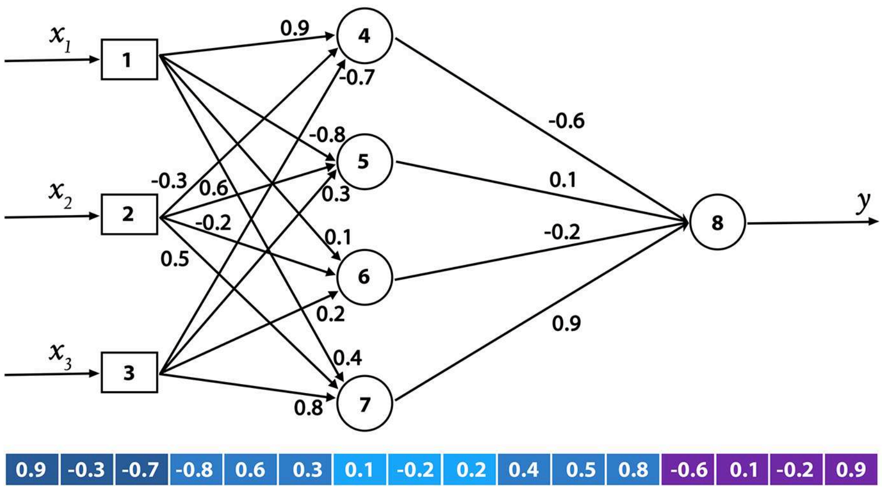

Figure 2. Weights of CNN as student status.

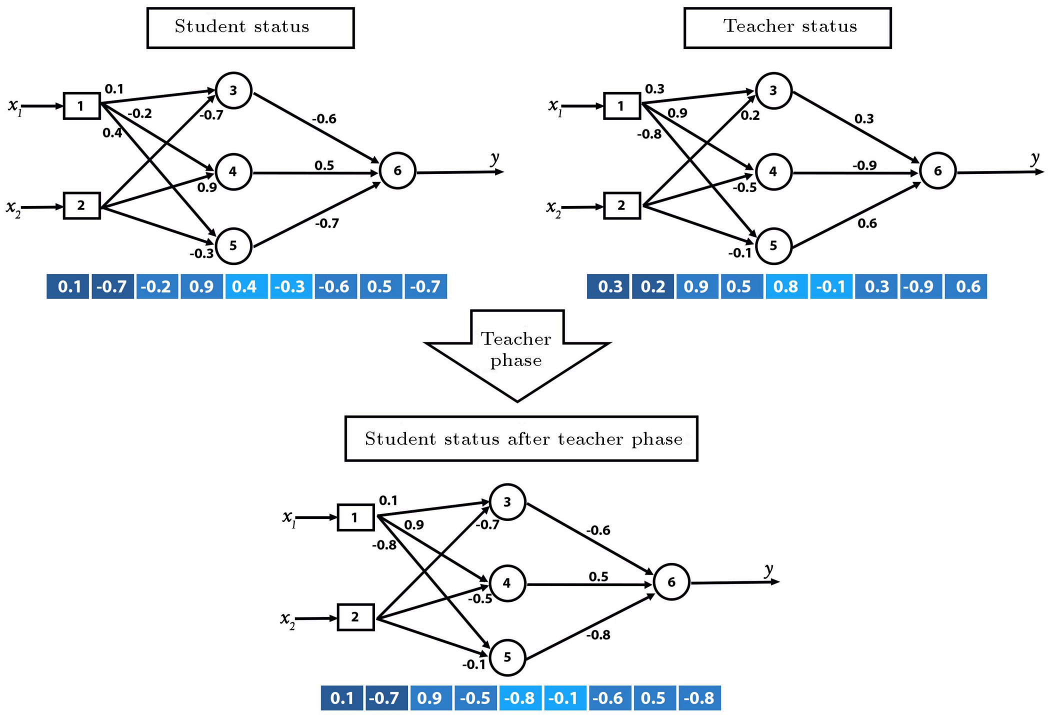

Figure 3. New status of the student after teacher phase.

student in terms of objective function value is selected as the teacher. Essentially, the teacher is a member of the students' population that has the lowest MSE;

- **Training phase (teacher phase):** In this step, the new status of each student is calculated based on the current situations of the student, teacher, and mean. The teacher causes the values of weights in the CNN to change. Figure 3 indicates the status of the student before and after the teacher phase. The new status of the student is evaluated by applying training inputs to it, calculating training outputs, and computing the MSE, which is considered as the new objective function value of the student. If the MSE of the student in the new status is better than the previous status, it is stored; otherwise, its storage is rejected;

- **Student phase:** In this step, the new status of each student, randomly selected based on the current situations of the student and another student, is calculated. The student phase causes the values of weights in the CNN to change. Figure 4 illustrates the student's status before and after the learning phase. The new status of the student is evaluated by

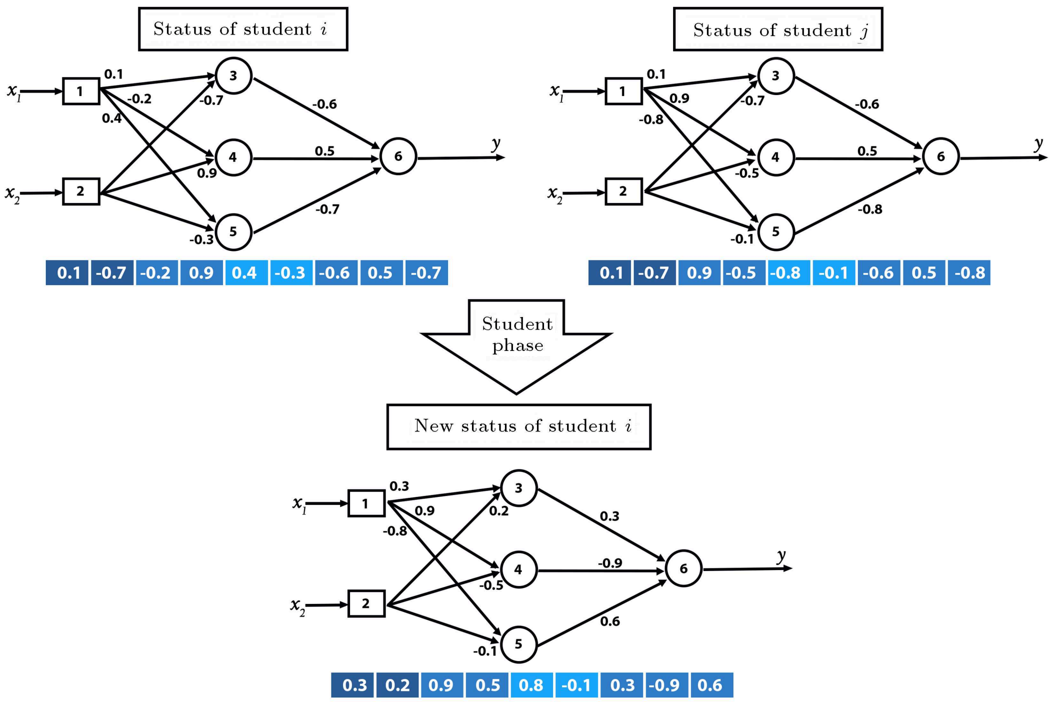

Figure 4. New status of the student after the learning phase in the student phase.

Table 1. General explanation of considered dataset.

Solar year	Solar month	Solar day	Dry temperature average (Celsius)	Absolute maximum temperature (Celsius)	Absolute minimum temperature (Celsius)	Average humidity	Maximum humidity	Minimum humidity
1395	1	1	23.2	26.2	20.7	68.5	60	79
1395	1	2	22.9	27.5	14.8	71.3	62	82
1395	1	3	21.4	23.4	21.4	81.9	74	87
1395	1	4	22.5	25.7	18.2	83.6	76	94
1395	1	5	22.6	26.5	19.0	65.9	59	78
1395	1	6	22.6	28.2	17.4	69.3	51	80
1395	1	7	22.1	27.6	15.6	72.3	60	82
1395	1	8	22.8	30.1	15.6	66.3	37	87

applying training inputs to it, calculating training outputs, and computing the MSE, which is considered as a new objective function value of the student. If the MSE of the student in the new status is better than the previous status, it is stored; otherwise, its storage is rejected.

The above operation is repeated to fulfill the finishing situation. The output of the optimization algorithm is based on the training and learning of the students, resulting in the best values of weights in the CNN, so that MSE is minimized [27].

4. Data analysis

The considered dataset is related to weather prediction for the country from 2016 onwards. This dataset has nine features. Table 1 is a part of the significant dataset, where in the range of data changes in each column is different. In this dataset, we used features such as year, solar day, solar month, and dry temperature average for predicting average humidity. A general explanation of this dataset is displayed in the following table.

4.1. Prediction of maximum humidity

First, an LSTM model with 52 blocks was applied to predict the maximum humidity. The results obtained by the LSTM model on the proposed data for predicting the maximum humidity are as follows.

Table 2 shows the Root Mean Square Error (RMSE) value for training and test data.

Table 2. The test and training RMSE.

	Train	Test
RMSE	9.42	8.64

Table 3. RMSE amount of CNNLSTM model for training and test data.

	Train	Test
RMSE	13.35	14.94

4.2. Combined model

In the combined proposed approach, the output of the CNN is given to an LSTM network which are considered as follows:

- A Convolution layer with 128 4*4 kernels;
- A-Max pooling layer with 2*2 kernel;
- A Convolution layer with 128 2*2 kernels;
- A Max pooling layer with 2*2 kernel;
- An LSTM network with 10 LSTM blocks;
- Dropout = 0.5;
- A layer connected with ten neurons;
- Last layer with one neuron for prediction.

The structure of the considerable CNNLSTM model is shown in Figure 5.

The results obtained by the CNNLSTM model on proposed data are shown in Figure 6.

Table 3 indicates the RMSE amount of the CNNLSTM model for training and test data.

Figure 7 indicates the prediction results of the CNNLSTM model on training and test data.

As observed, the GRU model obtained a lower RMSE compared to the other two models on the considerable dataset. On the other hand, the Bidirectional LSTM (BLSTM) model obtained a lower RMSE against the LSTM model (Figure 8).

4.3. Evaluation of model efficiency

To evaluate the efficiency of the models, we applied two criteria: Mean Absolute Error (MAE) and cosine proximity. Table 4 indicates the results of the proposed models based on these criteria.

The cosine proximity values are small across all models, but in terms of map, the GRU model has a lower value compared to the other models. The results for the prediction of maximum temperature are shown in Table 5.

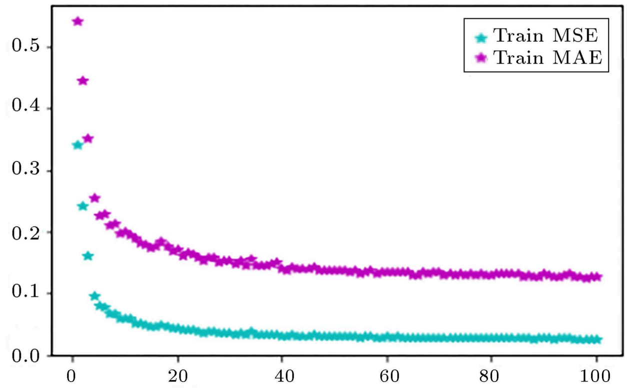

Figure 6. Diagram related to MSE and MAE values on training data.

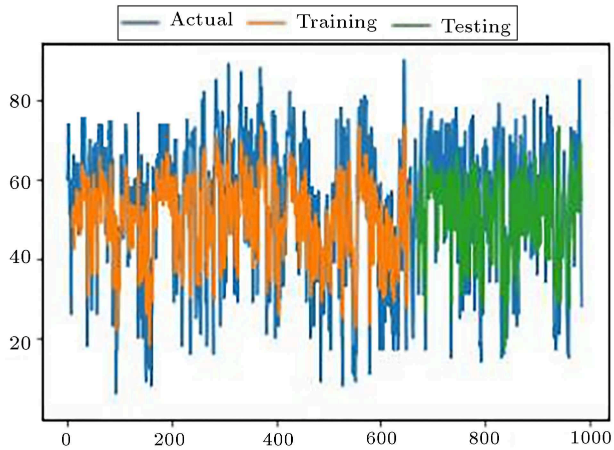

Figure 7. In this model, blue color shows actual data; yellow color indicates model prediction on training data, and green color shows model prediction on test data.

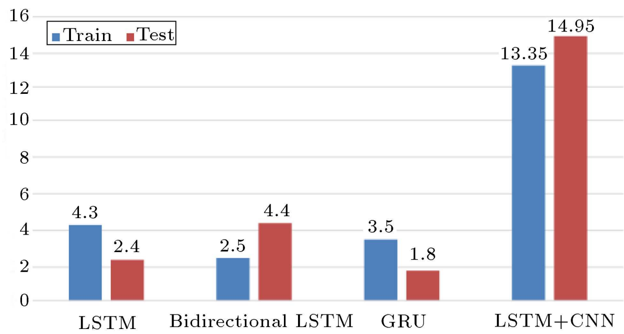

Figure 8. The prediction of model humidity with RMSE criterion.

For evaluating model efficiency, we applied two criteria: MAE and cosine proximity. Table 6 indicates the results of proposed models based on these two criteria.

Figure 5. The structure of considerable CNNLSTM model.

Table 4. Proposed models based on mean absolute percentage error and cosine proximity.

Model	MAE	Cosine proximity
LSTM	24.8340	$-1.0000e^{+00}$
BLSTM	21.6581	$-1.0000e^{+00}$
GRU	21.4872	$-1.0000e^{+00}$
CNN+LSMT	63.12	$-1.0000e^{+00}$

Table 5. Prediction of maximum temperature for training and test data.

Model	Train	Test
LSTM	2.31	2.31
BLSTM	2.12	2.17
GRU	1.98	2.08
CNN+LSMT	1.93	2.11

Table 6. Proposed models based on mean absolute percentage error and cosine proximity.

Model	MAE	Cosine proximity
LSTM	462428.4772	$-9.9846e^{-01}$
BLSTM	452009.3791	$-9.9846e^{-01}$
GRU	450656.1325	$-9.9846e^{-01}$
CNN+LSMT	632595.9525	$-9.9846e^{-01}$

Table 7. RMSE amount of different models for training and test data.

Model	Train	Test
LSTM	2.31	2.31
BLSTM	2.12	2.17
GRU	1.98	2.08
CNN+LSMT	1.93	2.11

4.4. The implementation of maximum temperature prediction

In the next part, the results of each model for predicting maximum temperature are brought. Table 7 shows the RMSE values of different models for training and test data.

5. Proposed method

5.1. Recurrent Neural Networks (RNN)

The present study considers decreasing prediction error for various types of time series and uncertainty in estimation parameters, improving the structure of the DNN, and increasing the response speed in the proposed neural network method. RNN are a type of forwarding network that adds a temporal dimension to the model, which is defined through edges in adjacent steps. In these networks, in each step, the status is related to current input and the previous step. If x^t is the current input and h^t is the previous status of the network which is obtained from the hidden node of the network, the network output is calculated by Eq. (1):

$$y^t = Soft\max\left(w^{yh}h^t + b_y\right), \tag{1}$$

and h^t is calculated through the following equation:

$$h^t = \partial\left(w^{hx}x^t + w^{hh}h^{(t-1)} + b_h\right). \tag{2}$$

In these equations, w^{yh}, w^{hh}, and w^{hx} are learnable weight matrices, and b_y and b_h are bias values that allow each neuron to learn the offsets. In these networks, during BP over a long time, the gradient will disappear. LSTM networks are a solution to prevent this problem.

Two following methods are used for the improvement of LSTM networks:

- **BLSTM networks:** LSTM networks, LSTM, are a particular type of RNN which can learn long-term dependencies. The purpose of designing LSTM networks was to solve the problem of long-term dependence. Remembering information for long-term intervals is a default and common behavior of LSTM networks, and their structures are such that they learn distant information, a feature which is hidden in their structure;
- **GRU networks:** GRU architecture, is presented to solve the deficiencies of traditional RNN such as gradient fading and the increased overhead in LSTM architecture. GRU is generally considered as a modified version of LSTM, as both architectures share the same design and obtain equally excellent results. This architecture uses the concepts of update gate and reset gate. These gates, represented by two vectors, determine whether information is conveyed to the output. What sets these gates apart is their ability to be trained until information related to past steps is kept without any change during different time steps.

Recurrent networks operate with GRU cells, such as the LSTM mechanism, and the only difference is that instead of three gates, they use two gates. As a result, we will have relatively faster performance and results that are almost the same as those of the LSTM model. Conceptually, a GRU network has two gates: r_t readjust and Z_t forgetting, which ensure that long-term dependencies of gradient fading do not occur. The values of these gates obtained by Eq. (3) as follows:

$$Z_t = \partial\left(W_z x_t + U_z h_{t-1} + b_z\right),$$

$$r_t = \partial\left(W_r x_t + U_r h_{t-1} + b_r\right). \tag{3}$$

Table 8 shows the RMSE value for training and test data.

Table 8. The test and training RMSE.

RMSE	Train	Test
	8.62	7.84

Table 9. The test and training RMSE.

RMSE	Train	Test
	9.44	8.59

This model obtained a lower RMSE value compared to the LSTM model. Additionally, its prediction results were more realistic than those of the LSTM model on both test and training data when compared against the significant data.

Back two-way neural networks (BLSTM) connect two hidden layers from different directions to the same output. Using this form of relative deep learning, the output layer can concurrently receive data from previous (back) and future (forth) modes. BLSTMs do not require input data; also, their input data are available in the previous status. Table 9 shows the RMSE value for training and test data.

6. Discussion and conclusion

In general, humidity is considered in two modes: absolute humidity and relative humidity. Absolute humidity is the amount of steam present in a unit volume of air. The amount of steam required for the saturation of a definite volume of air increases with temperature. Relative humidity is the present humidity amount in the air volume with a specific temperature relative to the maximum humidity that the air can have at the same temperature. Temperature is a physical and relative quantity that determines the amount of hotness and coldness and is measurable with a thermometer. When two objects have different temperatures, heat is transferred from the item with a higher temperature to the one with a lower temperature until they both reach an equal temperature.

The estimation of fault impedance and distance values due to the occurrence of a Single Line to Ground Fault (SLGF) and balanced fault in the underground distribution system is proposed. A single set of three-phase voltage and current waveforms is measured at the main substation. The Discrete Wavelet Transform (DWT) is used to extract the features from the measured voltage and current waveforms. Then, the correlation between the extracted features of voltage and current waveforms is obtained using cross-product analysis. Subsequently, an Artificial Neural Network (ANN) is utilized to estimate the fault impedance and distance values. A thorough investigation is carried out to analyze the effect of different types of ANN learning algorithms, as well as ANN parameters such as learning rate (lr), momentum constants (mc), and the number of neurons in a hidden layer.

In the present research, Section 4 examines the maximum prediction of humidity and temperature using various methods. The models used were samples of Recurrent Neural Networks (RNN). Among these networks, Gated Recurrent Unit (GRU), Bidirectional LSTM (BLSTM), Long Short-Term Memory (LSTM), and the proposed combined CNN+LST model were used. Among the criteria which were applied to evaluate the models are Root Mean Square Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), and Cosine Proximity.

Around, LSTM started to revolutionize speech recognition, outperforming traditional models in certain speech applications. In 2009, a Connectionist Temporal Classification (CTC)-trained LSTM network was the first RNN to win pattern recognition contests when it won several competitions in connected handwriting recognition. In 2014, the Chinese company Baidu used CTC-trained RNNs to break the 2S09 Switchboard Hub5'00 speech recognition dataset benchmark without using any traditional speech processing methods [28].

LSTM also improved large-vocabulary speech recognition and text-to-speech synthesis and was used in Google Android. In 2015, Google's speech recognition reportedly experienced a dramatic performance jump of 49% through CTC-trained LSTM [29].

The maximum temperature and humidity are predicted using GRU and LSTM. The RMSE method is used to compare the obtained values in these predictions, which is known as a sufficient amount. RMSE is an index that shows the effective amount of error. In other words, this index indicates the difference between the predicted values by the model and the actual values of a variable. This instrument for determining the model is selected as the model close to reality.

In the GRU network method, two gates are used instead of three for predictng the maximum humidity, and a relatively good speed of improvement is observed. However, the difference compared to the LSTM model is minimal, and they exhibit similar performance. In general, GRU networks produce more realistic results compared to other models, and back two-way networks produce better results on test data compared to LSTM networks.

References

1. Rao, R.V., Savsani, V.J., and Balic, J. "Teaching-learning-based optimization algorithm for unconstrained and constrained actual-parameter optimization problems", *Engineering Optimization*, **44**(12), pp. 1447–1462 (2012).
DOI: 10.1080/0305215X.2011.652103

2. Schmidt-Hieber, A.J. "Nonparametric regression using deep neural networks with ReLU activation function", *Annals of Statistics*, **48**(4), pp. 1875–1897 (2020). DOI: 10.1214/19-AOS1875

3. Gorgoglione, A., Gioia, A., and Iacobellis, V. "A framework for assessing modeling performance and effects of rainfall-catchment-drainage characteristics on nutrient urban runoff in poorly gauged watersheds", *Sustainability*, **11**(18), pp. 4933–4945 (2020). DOI: 10.3390/su11184933

4. Niu, D. and Dai, S. "A short-term load forecasting model with a modified particle swarm optimization algorithm and least squares support vector machine based on the denoising method of empirical mode decomposition and grey relational analysis", *Energies*, **10**(3), pp. 408–428 (2017). DOI: 10.3390/en10030408

5. Wax, M. and Adler, A. "Detection of the number of signals by signal subspace matching. ieee trans", *Signal Process*, **69**(1), pp. 973–985 (2021). DOI: 10.1109/TSP.2021.3053495

6. Yadav, S., Wajid, M., and Usman, M., *Support Vector Machine-Based Direction of Arrival Estimation with Uniform Linear Array*, In book: Advances in Computational Intelligence Techniques Publisher: Springer, pp. 253–264 (2020). DOI: 10.1007/978-981-15-2620-6

7. Choo, Y., Park, Y., and Seong, W. "Detection of direction-of-arrival in time domain using compressive time delay estimation with single and multiple measurements", *Sensors (Basel)*, **20**(18), pp. 5431–5442 (2020). DOI: 10.3390/s20185431

8. Park, C. and Lee, D. "Classification of respiratory states using spectrogram with convolutional neural network", *Applied Sciences*, **12**(4), pp. 1895–1906 (2022). DOI: 10.3390/app12041895

9. Ketkar, N. and Moolayil, J., *Convolutional Neural Networks. in Deep Learning With Python*, Apress: Berkeley, CA, USA. pp. 197–242 (2021). DOI: 10.1007/978-1-4842-5364-9_6

10. Wei, Y. and Jiang, Z. "Estimating parameters of structural models using neural networks", *USC Marshall School of Business Research Paper*, **22**(1), pp. 1–46 (2022). DOI: 10.2139/ssrn.3496098

11. Wood, A., Wood, R., and Charnley, M. "Through-the-wall radar detection using machine learning", *Results in Applied Mathematics*, **7**(1), pp. 100106–100114 (2020). DOI: 10.1016/j.rinam.2020.100106

12. Gregorczyk, M., Żórawski, P., Nowakowski, P., et al. "Sniffing detection based on network traffic probing and machine learning", In *IEEE Access*, **8**(1), pp. 149255–149269 (2020). DOI: 10.1109/ACCESS.2020.3016076

13. Cuntz, H. "Forest of synthetic pyramidal dendrites grown using cajal's laws of neuronal branching", *PLoS Computational Biology*, **6**(8), ev06.i08-17. (2010). DOI: 10.1371/image.pcbi.v06.i08

14. Bain, A., *Mind and Body. the Theories of Their Relation*, New York: d. Appleton and Company, James, the principles of psychology, New York: H. Holt and Company. (1873). URL: https://archive.org/details/mindbodytheories00bain

15. Brush, S.G. "History of the lenz-ising model", *Reviews of Modern Physics*, **39**(4), pp. 883–893 (1967). DOI: 10.1103/RevModPhys.39.883

16. Sherrington, C.S. "Experiments in examination of the peripheral distribution of the fibers of the posterior roots of some spinal nerves", *Proceedings of the Royal Society of London*, **190**(1), pp. 45–186 (1989). DOI: 10.1098/rstb.1898.0002

17. Li, Y., Lan, C., Xing, J., et al. "Online human action detection using joint classification-regression recurrent neural networks", In: *14th European Conference on Computer Vision - ECCV, Part VII*, Springer (2017). DOI: 10.48550/arXiv.1604.05633

18. Lahmiri, S. "Wavelet low- and high-frequency components as features for predicting stock prices with back propagation neural networks", *Journal of King Saud University - Computer and Information Sciences*, **26**(2), pp. 218–227 (2014). DOI: 10.1016/j.jksuci.2013.12.001

19. Ticknor, J.L. "A bayesian regularized artificial neural network for stock market forecasting", *Expert Systems with Applications*, **40**(14), pp. 5501–5506 (2013). DOI: 10.1016/j.eswa.2013.04.013

20. Kara, Y., Boyacioglu, M.A., and Baykan, O.K. "Predicting the direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange", *Journal of Expert Systems with Applications*, **38**(5), pp. 5311–5319 (2011). DOI: 10.1016/j.eswa.2010.10.027

21. Zhang, H.M., Wang, Z.B., Wu, Z.H., et al. "Real-time through-the-wall radar imaging under unknown wall characteristics using the least-squares support vector machines based method", *Journal of Applied Remote Sensing*, **10**(2), pp. 020501–0205011 (2021). DOI: 10.1117/1.JRS.10.020501

22. Kose, U. and Arslan, A. "Forecasting chaotic time series via anfis supported by vortex optimization algorithm: Applications on electroencephalogram time series", *Arab J Sci Eng.*, **42**(1), pp. 3103–3114 (2017). DOI: 10.1007/s13369-016-2279-z

23. Ma, X., Jin, Y., and Dong, Q. "A generalized dynamic fuzzy neural network based on singular spectrum analysis optimized by brain storm optimization for short-term wind speed forecasting", *Applied Soft Computing*, **54**(1), pp. 296–312 (2017). DOI: 10.1016/j.asoc.2017.01.033

24. Zhu, W., Lan, C., Xing, J., et al. "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks", *AAAI Conference on Artificial Intelligence* (2016) DOI: 10.48550/arXiv.1603.07772

25. Zhang, S., Liu, X., and Xiao, J. "On geometric features for skeleton-based action recognition using multilayer lstm networks", *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2017). DOI: 10.1109/TMM.2018.2802648

26. Qiu, X., Ren, Y., Suganthan, P.N., et al. "Empirical mode decomposition based ensemble deep learning for load demand time series forecasting", *Applied Soft Computing*, **54**(1), pp. 246–255 (2017). DOI: 10.1016/j.asoc.2017.01.015

27. Tarigan, J., Diedan, R., and Suryana, Y. "Plate recognition using backpropagation neural network and genetic algorithm", *Procedia Computer Science*, **116**(1), pp. 365–372 (2017). DOI: 10.1016/j.procs.2017.10.068

28. Hannun, A.Y., Case, C., Casper, J., et al. "Deep speech: Scaling up end-to-end speech recognition", *ArXiv.* (2014). DOI: 10.48550/arXiv.1412.5567

29. Sak, H., Senior, A.W., Rao, K., et al. "Learning acoustic frame labeling for speech recognition with recurrent neural networks", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2015). DOI: 10.1109/ICASSP.2015.7178778

Biographies

Hamideh Rezaei Nezhad is currently a PhD student and a member of the academic staff at Qeshm Islamic Azad University. She received her BSc degree in Software Engineering from Baft Islamic Azad University and her MSc degree in Software Engineering from Qeshm International University. She is currently pursuing a doctoral degree in computer science. Her main specialties include estimation and forecasting, deep learning, processing complex systems, data mining, software engineering, and artificial intelligence.

Farshid Keynia is currently affiliated with the Department of Energy Management and Optimization at the Institute of Science and High Technology and Environmental Sciences, Graduate University of Advanced Technology, Kerman, Iran, He received his BSc degree in Electrical Engineering from Shahid Bahonar University, Kerman, Iran, in 1996, and his MSc degree in Electrical Engineering from Semnan University, Semnan, Iran, in 2001. He is completed his PhD in the Electrical Engineering Department of Semnan University, Semnan, Iran, in 2010. His research focuses on electricity market features, short-term and mid-term price and load forecasting in deregulated electricity markets, feature selection, classification algorithms, and metaheuristic multi-optimization methods.

Amir Sabbagh Molahosseini Amir Sabbagh Molahosseini received his BSc degree from Shahid Bahonar University of Kerman, Iran in 2005, and his MSc and PhD degrees (Hons.) in computer engineering from the Science and Research Branch of Azad University, Tehran, Iran, in 2007 and 2010, respectively. He is currently with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, UK. His research interests are Computer Arithmetic/Architecture and Approximate/Transprecise Computing. HYPERLINK "mailto:a.sabbaghmolahosseini@qub.ac.uk; a.sabbaghmolahosseini@qub.ac.uk"