Received August 13, 2021, accepted August 26, 2021, date of publication September 14, 2021,
date of current version September 30, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3112666

Electricity Consumption Forecasting Using Gated-FCN With Ensemble Strategy

AQDAS NAZ[1], NADEEM JAVAID[1], (Senior Member, IEEE), MUHAMMAD ASIF[1], MUHAMMAD UMAR JAVED[1], ABRAR AHMED[2], SARDAR MUHAMMAD GULFAM[2], MUHAMMAD SHAFIQ[3], AND JIN-GHOO CHOI[3]

[1]Department of Computer Science, COMSATS University Islamabad, Islamabad 44000, Pakistan
[2]Department of Electrical and Computer Engineering, COMSATS University Islamabad, Islamabad 44000, Pakistan
[3]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, South Korea

Corresponding authors: Nadeem Javaid (nadeemjavaidqau@gmail.com), Muhammad Shafiq (shafiq@ynu.ac.kr), and Jin-Ghoo Choi (jchoi@yu.ac.kr)

This work was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2021-2016-0-00313) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

ABSTRACT Accurate electricity consumption forecasting in the power grids ensures efficient generation and distribution of electricity. Keeping this in mind, the paper introduces a novel deep learning model, termed Gated-FCN, for short-term load forecasting. The key idea is to introduce an automated feature selection and deep learning model for forecasting. The model includes an eight-layered Fully Convolutional Network (FCN-8) in which the hand-crafted feature selection that requires expert domain knowledge is avoided. Furthermore, the model also reduces noise as it learns internal dependencies and the correlation of the time series. Enhanced Bidirectional Gated Recurrent Unit (EBiGRU) is used in combination with FCN-8 to learn long-term temporal dependencies of the time series. Moreover, a weighted averaging mechanism of multiple snapshot models is adopted in the proposed model to assign optimized weights to BiGRU. At the end of FCN-8 and BiGRU, a fully connected dense layer is used that gives final prediction results. Gated-FCN is an end-to-end forecasting model that does not require any other model for enhancing its forecasting efficiency. Different activation functions are initially analyzed to determine how the proposed model learns complex patterns from the time series data. Later, the activation function having the best accuracy is used for forecasting. The proposed model extracts both spatial and temporal features from the data. Furthermore, this paper also provides predictive and exploratory data analyses to assist policymakers in making optimal decisions regarding power production and dispatch. In order to demonstrate the applicability of the proposed technique, the simulations are performed using nine years' load consumption data taken from Independent System Operators New England (ISO-NE). The comparison with five state-of-the-art techniques is also provided to prove the fact that Gated-FCN gives the best forecasting accuracy as compared to other benchmark techniques in terms of two performance metrics: Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE).

INDEX TERMS Deep learning forecasting technique, enhanced BiGRU, electricity consumption forecasting, gated FCN, smart grid, weighted averaging technique.

NOMENCLATURE

ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
BiGRU	Bidirectional Gated Recurrent Unit
BiLSTM	Bidirectional Long Short Term Memory
CNN	Convolutional Neural Network
DANN	Deep Artificial Neural Network
DLSTM	Dilated Long Short Term Memory
DR	Demand Response
EBiGRU	Enhanced Bidirectional Gated Recurrent Unit
ELU	Exponential Linear Unit
FCN-8	Fully Convolutional Network-8

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

GA	Genetic Algorithm
Gated FCN	Gated Fully Convolutional Network
GRU	Gated Recurrent Unit
ISO-NE	Independent System Operators New England
LSTM	Long Short Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MSE	Mean Square Error
MLP	Multi Layer Perceptron
MLR	Multi Linear Regression
PReLU	Parametric Rectified Linear Unit
RBM	Restricted Boltzmann Machine
ReLU	Rectified Linear Unit
RES	Renewable Energy Source
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
SANN	Shallow Artificial Neural Network
SELU	Scaled Exponential Linear Unit
SMI	Smart Metering Infrastructure
SMA	Simple Moving Average
SVM	Support Vector Machine

I. INTRODUCTION

The importance of load forecasting cannot be overlooked in the modern power management system. Electricity infrastructure planning and power system operations are supported by long-term, mid-term and short-term load forecasting. In the last couple of decades, enormous efforts have been made to ensure the power systems' sustainability. The penetration of Renewable Energy Sources (RESs) in the power systems is also on the rise to fulfill the increasing demand for energy. The uncertainty in the power systems increases due to the intermittent nature of RESs (inconsistent availability of wind and solar resources) [1]. Moreover, the Electricity Consumption (EC) pattern also shows variations. Both these factors directly affect the balance of electricity demand and supply. Therefore, accurate short-term EC load forecasting is necessary. It allows the power distributors to plan the operations of energy generation from different sources while maintaining the reliability and sustainability of the power system. In the case of surplus energy generation from RES, the power distributors can use EC load forecasting information to decide how much energy is generated by RESs should be stored for later use. On the other hand, in the case of energy deficit, they can plan the fuel-based energy generation to meet the additional demand. Efficient load forecasting also helps in intelligent usage of storage devices and load shaving through Demand Response (DR) system. Thus, load forecasting can be used by the utilities for DR programs to handle power deficiency and to ensure load balancing at an optimal level.

The authors in [2] introduce an ensemble and deep learning-based short-term forecasting model. The proposed approach is the combination of Light Gradient Boosting Machine (LGBM), Extreme Gradient Boosting (XGB) and Multi-Layer Perceptron MLP) models. The MLP first learns the hidden features' representation from the electricity load profile and then ensemble boosting techniques are used to efficiently deal with the huge electricity load dataset. The ensemble techniques perform well as compared to traditional machine learning techniques because of the combined integration of several weak learners. The result is finalized by considering the outcomes of all weak learners. That is the reason for adopting the boosting machine in the proposed methodology for efficient short-term forecasting of EC load profile. However, the proposed scheme lacks while capturing the temporal dependencies from the electricity load data because the MLP or ensemble models are unable to maintain the context of electricity load profile for a long period. Recently, a KNN and deep learning-based hybrid model is introduced in [3] for short-term load forecasting. The KNN is exploited to capture the uncertainty and fluctuations from the electricity load consumption data. Meanwhile, the deep belief network is employed to obtain the prediction interval. However, the proposed scheme performs poorly while forecasting electricity load consumption because the KNN is a lazy learner. It does not capture the temporal correlated features from the long-term electricity load profile properly. Furthermore, the study of [4] presents an enhanced CNN by utilizing the neuroevolution algorithm as a hyperparameter tuning algorithm. However, the standalone CNN does not accurately forecast the electricity load profile. It does not have enough memory to maintain and store the long-term context of the electricity load profile. After critically reviewing the recent literature, it is observed that there is still room for an accurate EC load forecasting model. Keeping the above concerns in view, we propose a hybrid deep learning model, which is the combination of FCN and BiGRU. The problems of long-term dependency and high targeting features are resolved by BiGRU and FCN, respectively. The BiGRU intelligently captures the temporal correlation from the electricity load data by using the memory gates and then forecasts an accurate short-term load profile.

The conventional metering infrastructure is gradually being replaced by Smart Metering Infrastructure (SMI). This massive change drives the utility to rely on load forecasting for power management and operations. Several forecasting techniques are introduced in the literature to efficiently perform electricity load forecasting. Artificial Neural Network (ANN) was firstly introduced in 1943. The concept was based on biological neurons. ANN-based learning techniques include Shallow ANNs (SANNs) and Deep ANNs (DANNs). Besides, the Convolutional Neural network (CNN) is also used for classification purposes. It was introduced for the first time in 1980. It is based on the concept of neurons that are connected together in the form of a multi-layer hierarchy [5]. Deep learning techniques provide better results than shallow learning techniques because the latter face training and gradient diminishing issues. Thus, deep learning techniques are being used in many fields, such as computer vision, speech recognition and natural language processing [6]. A DANN stacks multiple layers of neural networks in different sequences and uses stochastic optimization methods to perform the machine learning tasks. Moreover, it does not require hand-crafted features like machine learning techniques. All layers in the network are jointly trained, which is one of the challenges being faced by deep learning models [7]. Many variants of DANN have been proposed in the literature for electricity load forecasting, such as Restricted Boltzmann Machine (RBM) [8] and Multi-Layer Perceptron (MLP) [9]. However, the increase in the number of layers increases the complexity of the network in terms of execution time and memory requirement.

Long Short Term Memory (LSTM) is an advanced and powerful version of Recurrent Neural Network (RNN). It consists of three memory gates that save historical information [10]. These memory gates include input, output and forget gates. Several flavors of LSTM have been proposed in the literature, such as Dilated LSTM (DLSTM) and Bidirectional LSTM (BiLSTM). CNN learns the correlation between spatial features. Whereas, LSTM learns temporal dependencies [5]. Researchers have combined the benefits of both models and proposed an end-to-end forecasting model. Convolutional LSTM was firstly used in 2015 for image recognition [11]. The prediction is made by learning the movements of pixels from previous frames. Various studies have been conducted in different fields to perform both feature extraction and forecasting tasks, simultaneously. In natural language processing, emotions are studied using text as an input [12]. In speech recognition, various voice search tasks are performed using CNN and LSTM. Combined CNN and LSTM model shows flexibility and gives good results despite the presence of noise. These forecasting models are also in the limelight in the image recognition field. In video processing, a deep learning model using CNN and BiLSTM is proposed to detect various human gestures in video frames [13]. Similarly, in the medical domain, arrhythmia disease is detected in ECG using a CNN-LSTM model [14].

In [1], the authors use LSTM for learning temporal trends in time series data. However, the correlation of exogenous and dependent variables plays a vital role in learning the time series data, which is not considered in this paper. The fully Convolutional Network (FCN) learns the correlation between exogenous variables, extracts important features and passes them as inputs for forecasting future EC. Hence, manual crafting of features is avoided while filtering the data as it is labor extensive. Furthermore, LSTM uses three gates: input, output and forget, which increase the execution time of the proposed model. To tackle the aforementioned issue, Gated Recurrent Unit (GRU) was proposed in 2014, which is an advanced version of LSTM. GRU has only two gates. It combines the input and the forget gate of LSTM into an update gate. The remaining mechanism does not change and is the same as that of LSTM [15]. It takes less memory as compared to LSTM, yet produces good results. GRU is implemented on statistical machine translation. Many different flavors of GRU have been introduced in literature [16]. Bidirectional GRU (BiGRU) is also one of the variants. It learns temporal dependencies in the data both from the past and future.

In the proposed work, EC is predicted using a combined FCN-8 and Enhanced BiGRU (EBiGRU) model, named Gated-FCN. FCN-8 acts as an encoder and reduces data dimensionality. It extracts spatial features from the data to decrease its complexity. Moreover, it neither requires manual handcrafting of the exogenous features nor the preprocessing of data. EBiGRU is effective in modeling the temporal information hidden in irregular data trends and learning the long-term historical dependencies. EBiGRU uses an ensemble strategy to increase the generalization capability of the model. Gated-FCN is different in its structure because it is an end-to-end forecasting model that learns both the correlation between exogenous variables and the long-term dependencies in the data. Independent System Operators New England (ISO-NE) dataset [47] is used to forecast short-term electricity consumption data. The simulation results prove that the proposed model outperforms the benchmark deep learning and shallow machine learning models.

The remaining paper is organized as follows. Section II presents the related work of the EC forecasting models. Section III briefly explains the working of BiGRU and FCN-8, and presents the complete methodology of the proposed model. Section IV elaborates simulation results. Section V concludes the paper.

II. RELATED WORK

Many studies are conducted in literature for EC forecasting, which are based on statistical, machine learning and deep learning models. There are several techniques based on these models that are proposed by researchers in the course of time. These techniques include Auto-Regressive Integrated Moving Average (ARIMA) [17], LR [18], SVM [19], ANN [20], sequence to sequence learning [21], DANN [22], etc. Table 1 presents the main findings of literature review in terms of proposed schemes, contributions, performance metrics, results and limitations. According to the statistics provided in [23] regarding the application of these forecasting techniques in research, ANN is being used up to 47%, SVM 25%, Decision Tree (DT) 4% and the rest of the techniques are employed up to 25% in the literature. The forecasting of residential electricity load is performed using ANN in [24]. However, the effectiveness of ANN increases as per the availability of different exogenous variables like weather data, temperature, wind speed and humidity. The nonavailability of dependent variables adversely affects the performance of ANN. RNN resolves this issue with the help of BP in the network. It yields promising results while forecasting sequence-based time series. However, long-term dependencies are hard to learn by traditional RNN as it faces vanishing gradient issues due to time depth and single hidden layer. In order to overcome these issues, the LSTM forecasting technique was proposed in 1997 [10]. It combines short-term memory with long-term memory by incorporating three memory gates, which save the long-term dependencies

TABLE 1. Literature review findings.

Proposed Scheme & Contributions	Metrics & Results	Limitations
A deep belief network is exploited for short term EC load forecasting while KNN is utilized for capturing the uncertainties and fluctuations from EC load profile [3]	Time horizon: 30 minutes; MAPE = 5.91%; RMSE = 6.89%	KNN is a lazy learner as it does not capture temporal correlated features from EC load consumption
The aim of [17] is to fulfill the energy demand of Queensland, Australia's second largest state, by providing a reliable EC load forecasting model termed as ARIMA	Pearson Product-Moment Correlation coefficient (r) = 0.990; Willmott's Index (WI) = 0.995; MAE = 55.915 MW	Long term dependency issues
Partial Functional Linear Regression Model (PFLRM) is proposed for forecasting the day-ahead EC load [18]	RMSE = 59.1704; MAD = 40.6277; MAPE = 0.1134	PFLRM lacks in dealing with highly correlated time series EC load data
A multi-input and multi-output based LSTM model is trained to forecast electricity load consumption according to season, day and interval [22]	RMSE train = 10.67%; RMSE test = 13.16%	Slow convergence and high computational overhead due to the excessive use of memory gates
In [29],[30], deep LSTM is used to perform efficient short term load forecasting by capturing the strongly correlated patterns from historical EC load data	MAPE = 3.52%; RMSE = 4.56%	Increased computational overhead due to excessive computations in memory gates
MLR is used for forecasting the electricity load data for South Sulawesi Electrical System [39]	MAPE in rainy season = 4.34%; MAPE in dry season = 3.52%	MLR lacks in capturing the long term temporal correlations
In [40], IEMD is used to decompose the time series EC signal and then passes it to a deep belief network. A correlation analysis is also performed using T-Copula for measuring the effect of external factors	VIC dataset (Jan-Mar) MAPE = 6.96; RMSE = 598.76	Computationally costly because of sequential integration of models
In [41], different ensemble ML models are exploited for efficient electricity load forecasting, where GBRT outperforms the other models	RMSE = 270.6 MW; MAPE = 1.32%	Curse of dimensionality
In [42], Bayesian network is designed	Performance with normal distribution network using 18 neurons: MAPE = 0.7083%; RMSE = 16.9378 MW	Uncertainties in results due to the existence of probability

in them. In [25], LSTM is combined with GA in order to optimize the model's parameters. The comparison between different deep learning and machine learning models is performed in [26], which proves that the performance of deep learning models is reasonably better as compared to the machine learning models. DANN is used as the forecasting technique in [27]. It is a multichannel deep convolution network, which uses variables' data to detect latent features. The extracted latent features are fed to MLP for forecasting the time series data. However, MLP performs poorly when detecting temporal dependencies. In [28] and [29], deep LSTM is used in order to perform forecasting.

Various variants of LSTM are proposed in the literature that include DLSTM. It uses skip connections that are extracted from the concept of ResNet [6]. Moreover, DLSTM enhances the efficiency of the network by reducing the vanishing gradient issue [30]. BiLSTM is used in literature to learn two-way dependencies of the data: past and future [31]. GRU is an advanced version of LSTM that combines the input gate and forget gate of the network into one gate, termed as an update gate [15]. Multi GRU is also a variant of GRU that is used to optimize the electricity dispatch plan [16]. In [32], GRU is used in order to forecast PV power generation. Apart from time series sequence data, there are other high dimensional information as well like spatiotemporal matrix that exists in the time series, which cannot be learned by GRU. Therefore, it is required to add such features in GRU that can optimally learn the high dimensional temporal data. In deep learning methods, feature extraction is ideally performed by CNN. It is widely used in the field of image and pattern recognition. It also learns local features that are based on a strong relationship between nearby points [33]. Researchers in the fields of image recognition, speech recognition and emotion recognition frequently use CNN. It is an end-to-end learning model that simultaneously covers spatial and temporal trends [34]. Moreover, CNN-LSTM is used in speech recognition to learn global and local emotion-based features. Multiple blocks are used in the network, comprising of convolution block and pooling block that learn the local correlation in the data. LSTM learns long-term dependencies of the extracted local features with incredible learning precision [35]. CNN-LSTM is also used in the field of natural language processing by taking the text as an input. The authors in [12] combine three deep learning models that include CNN, LSTM and DNN. These models provide the best results in emotions' recognition from voice [36]. Similarly, the authors perform action recognition from the video, which involves frame extraction for data processing. The proposed CNN-LSTM model is used for learning the users' actions [13]. Both image recognition from videos and emotion recognition from voice are performed using end-to-end forecasting frameworks that provide high forecasting precision. The forecasting of energy consumption time series is performed to learn temporal trends without handling high dimensional features. Furthermore, feature extraction plays an important role in noise and dimensionality reduction. It also reduces the computational time of the forecasting model and improves forecasting accuracy. In [37], the authors use a Multiple Linear Regression (MLR)

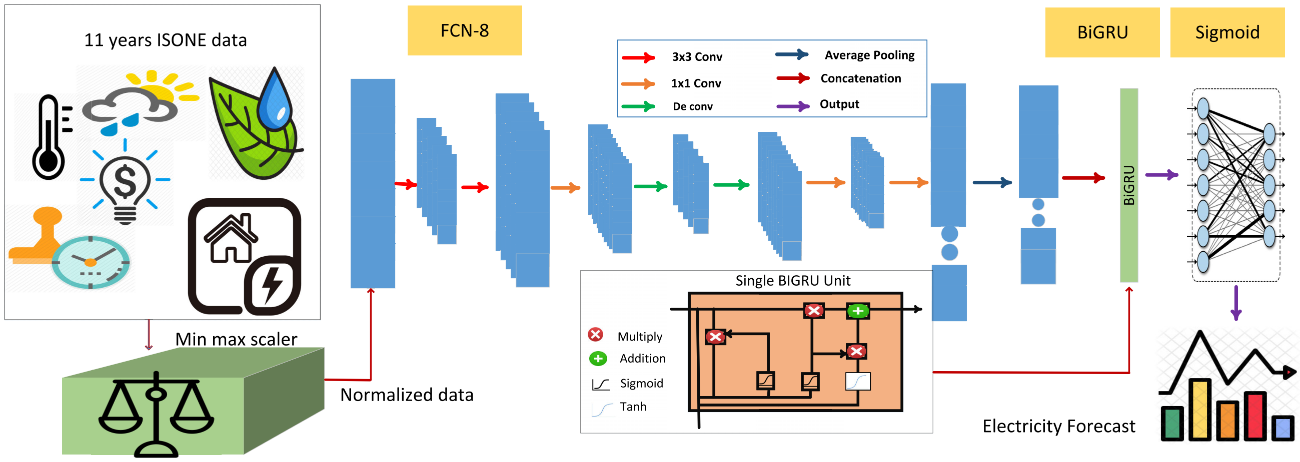


FIGURE 1. **Gated-FCN architecture.**

model for forecasting the electricity load data for South Sulawesi Electrical System. The proposed model performs efficient forecasting and provides an optimal electricity load forecast for rainy and dry seasons with a minimum MAPE of 4.34% and 3.52%, respectively. However, the EC load data contains long-term periodicity and fluctuations, which limit the performance of MLR in some cases. Similarly, in [4], an enhanced deep learning architecture is proposed for short term electricity load forecasting. In the proposed work, a deep CNN is built for extracting the complex non-linear patterns from the historical EC load profile and providing an optimal short-term electricity load forecast. The performance of CNN is enhanced by employing an evolutionary Enhanced Grey Wolf Optimizer (EGWO) technique for parameter tuning. The simulation results depict that the proposed model outperforms the other schemes. However, due to the presence of strong temporal correlation in the EC load profile, the CNN lacks in providing an optimal forecast because of short-term dependency issues. In [38], it is argued that some external factors like weekdays or weekends, weather conditions, seasonal changes, etc., also have an effect on short-term electricity load forecasting. In this regard, a novel model equipped with a signal decomposing technique, named Improved Empirical Mode Decomposition (IEMD), is proposed. IEMD decomposes the high dimensional time series signal into low-frequency components for better interpretation of EC load data. Moreover, for handling the effects of external variables, a correlation analysis is performed using T-Copula. Finally, the outcomes of both IEMD and T-Copula are passed as input to a deep belief network for increasing the model's efficiency. The proposed model outperforms the other baseline models in terms of achieving minimum MAPE and RMSE. The authors in [39] present a study for short term electricity load forecast. Four different statistical and ensemble learning models are used to forecast the day-ahead electricity load consumption. The simulation results prove that the ensemble model, named Gradient Boosting Regression Trees (GBRT), outperforms other methods in terms of load forecasting. In [40], a Bayesian neural network-based model is presented. The Bayesian network model is a type of probabilistic model, which tries to map conditional dependencies. The simulation results demonstrate that the Bayesian model provides an optimal load forecast as compared to other baseline models.

In [14], the authors extract spatial and temporal features in the time series using an end-to-end forecasting model. Multidimensional data consists of complex features, which affect the time series trends. CNN is used to learn complex spatial features. Whereas, LSTM is used to learn temporal-based features of the data. There are several irregular trends and long-term dependencies in the EC data that can easily be learned using LSTM. However, there are certain drawbacks of LSTM, which are tackled by a variant of LSTM, known as GRU. As already mentioned, GRU combines the input gate and the hidden gate, which reduces the computational time without compromising its efficacy. The variant of GRU that is used in the proposed work is termed EBiGRU. It learns long-term dependencies of the data both from the past and future. An eight-layered FCN (FCN-8) is used in the proposed work to extract spatial features from the time series data. Details of the proposed end-to-end forecasting model, Gated-FCN, are given in Section III.

III. PROPOSED DEEP LEARNING MODEL

The novel architecture that is proposed in this paper is based on two levels. At the first level, multiple exogenous features are taken as inputs without applying any manual feature crafting technique. The processing of data is performed using FCN-8 in order to extract important spatial correlation of the features and reduce noise. The output is flattened and is given as an input to BiGRU for the regression purpose. A multi-model ensemble technique is used to perform averaging of multiple snapshot models of BiGRU in order to increase the robustness and generalization capability of the model. The proposed architecture is illustrated in Fig. 1.

A. FULLY CONVOLUTIONAL NETWORK-8

CNN has been used in literature for pattern recognition in the image processing field [41]. It uses weights' sharing strategy while learning the time series data. Therefore, it does not require classic feature crafting methodologies or manual feature crafting capabilities. It quickly learns complex features from the data. Recently, CNN is used in electricity-related problems, such as forecasting RES generation. There are primarily three layers that are used in CNN: convolution, pooling and activation. The convolution layer takes a number of output activations of the previous layer as input. It has the size of filter and stride. In the proposed CNN, there are 8 convolution layers, which is why it is termed FCN-8. Filter, which is taken in the network is of small size as compared to the actual size of the time series window. It moves on small parts instead of the entire activation map. FCN-8 extracts the feature's correlation in the load consumption time series and reduces noise. The details of the layers involved in CNN are provided below.

1) CONVOLUTION LAYER

In convolution operation, a filter is applied to the input vectors and a feature map is obtained. φ shows the activation function. $\varphi(W^n \otimes x + b^n)$ is the convolution operation, W represents the weighting factor and n is the number of filters. x depicts the input vector and b shows the bias vector. There are multiple activation functions that are being used in literature, such as Rectified Linear Unit (ReLU), Leaky ReLU, Parametric ReLU (PReLU) [42], Exponential Linear Unit (ELU) and Scaled ELU (SELU) [43]. These are explained below.

- *ReLU*: it is used in deep learning models to solve numerous real-world problems. Due to its efficient learning power, it becomes one of the frequently used activation functions, which reduces both the vanishing gradient issue and the overfitting problem in the network. It is defined in Equation 1, [42].

$$f(x) = max(0, x), \tag{1}$$

  where x denotes the input data and $max()$ picks the maximum value. There are various issues that ReLU may encounter, such as dying ReLU. It happens when the output of ReLU is always zero regardless of the input because weights or biases are updated to be negative during the training phase. It causes the ReLU to be dead, which makes the learning of the model slow. In order to solve its inherited issues, many other activation functions have been introduced over the course of time that include Leaky ReLU, PReLU, ELU and SELU, as shown in Fig. 2.
- *Leaky ReLU*: it allows a small and positive gradient when unit is not active. In ReLU, neurons are not activated, which leads to poor results. In order to alleviate the problem of ReLU, non negative gradient is used, which improves the network accuracy. It is defined in Equation 2, [42].

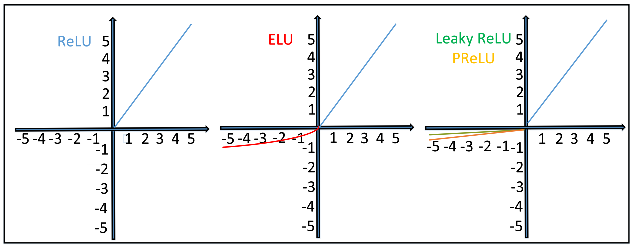


FIGURE 2. Activation functions.

$$f(x) = \begin{cases} x, & \text{if } x \geq 0, \\ 0.01x, & \text{otherwise.} \end{cases} \tag{2}$$

- *PReLU*: it also helps in reducing dying neuron issue by taking substitute of non zero gradient from the parameter a like Leaky ReLU. It takes a variable rather than taking the constant value of 0.01, as shown in Equation 3, [42].

$$f(x) = \begin{cases} x, & \text{if } x \geq 0, \\ ax, & \text{otherwise.} \end{cases} \tag{3}$$

- *ELU*: it is also based on ReLU. It has taken the good part of ReLU while reducing dying neuron problem. It improves the learning curve and solves the vanishing gradient problem. It is given in Equation 4, [43].

$$f(x) = \begin{cases} a(e^x - 1), & \text{if } x \geq 0, \\ x, & \text{otherwise,} \end{cases} \tag{4}$$

  where a is the scaling parameter and e denotes the exponential function, which calculates the small decay in input x.
- *SELU*: It is similar to ELU. The only difference is that it uses two parameters in order to reduce vanishing gradient issue, which improves the learning speed of the model. It is given in Equation 5, [43].

$$f(x) = \begin{cases} \lambda(ae^x - 1), & \text{if } x \geq 0, \\ x, & \text{otherwise,} \end{cases} \tag{5}$$

  where the value of α is 1.6733 and λ shows a stochastic variable, whose value is 1.0507.

2) POOLING LAYER

In FCN-8, the convolution layer is followed by a pooling layer. It is interposed between convolution layers after nonlinear activation functions. It is applied on the convolution layer and its main purpose is to reduce the size of data and avoid overfitting. Max(.) operation is used in order to spatially reduce the data, as illustrated in Fig. 3.

B. ENHANCED BIDIRECTIONAL GATED RECURRENT UNIT

ANNs are broadly used in literature to solve nonlinear real-world problems. RNN is used to learn the forward and

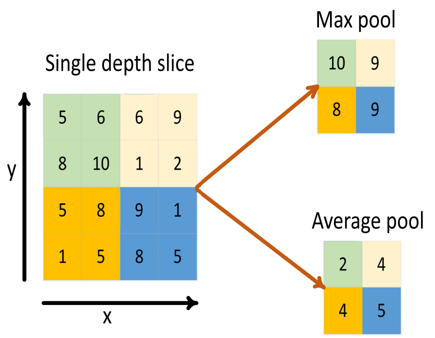


FIGURE 3. Max pooling and average pooling.

backward direction of the time series. BiGRU is one of the advanced versions of RNN and is used in the proposed work. It consists of two layers. The first layer learns the forward temporal sequence while the second layer learns the backward sequences. Both hidden layers' feature maps are concatenated and given as an input to the fully connected layer [44]. BiGRU encodes the temporal sequence of the features extracted through FCN-8. The previous layer dependencies are based upon the update gates, which control the flow of information from the previous hidden layer to the current hidden layer. On the other hand, the reset gate controls the information that is required to be neglected and not to be passed to the next hidden layer. The architecture of BiGRU is illustrated in Fig. 4.

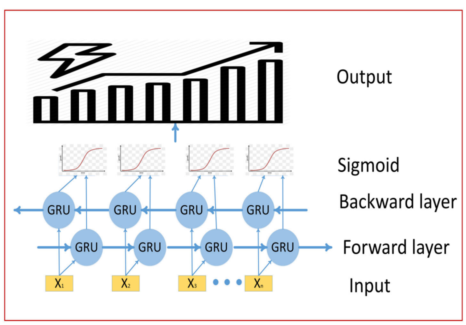


FIGURE 4. BiGRU architecture.

In machine learning, ensembles of different models increase the generalization capability of the ensemble model as compared to the individual model [45]. In the proposed model, a multimodel ensemble strategy is used to increase the generalization of the proposed Gated-FCN model, as shown in Fig. 5. In EBiGRU, an ensemble of 10 snapshot models is used. Furthermore, the Adaptive Moment Estimation (ADAM) optimizer is used to optimize EBiGRU because its performance is up to the mark in terms of efficacy. The algorithm performs adaptive learning by first and second moments estimation of the gradients. The speed of convergence is proven to be the fastest as compared to other optimizers. It is computationally fast and less memory-hungry [22]. Furthermore, the cycling learning rate in the case

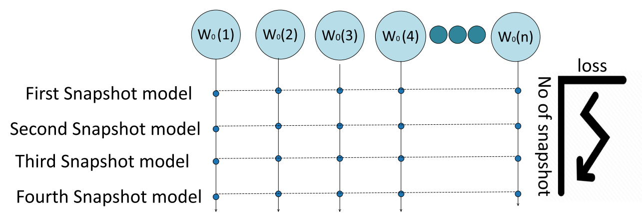


FIGURE 5. Ensemble strategy illustration - The number of snapshot models are selected on the basis of minimum loss.

of Stochastic Gradient Descent (SGD) adversely affects the performance of the forecasting model. Therefore, ADAM is used in our work as an optimizer, which iteratively adapts the learning rate. Moreover, it does not need to set the learning rate initially. The weights of the model, given as $W_0(1), W_0(2), W_0(3), \ldots, W_0(n)$ are trained multiple times. EBiGRU keeps the rest of its parameters the same, such as hidden layers and the number of neurons. Each snapshot model trains the weights differently because of the initial random values of the weights. The final weights of the models are then averaged in order to increase the robustness of the model.

C. CASCADED FCN-8 AND BiGRU

Gated-FCN is a combination of FCN-8 and EBiGRU. The core reason for using FCN-8 as an encoder is to extract spatial features and reduce noise from the multivariate dataset. Subsequently, EBiGRU is used in order to learn encoded local temporal features and long-term dependencies. It learns in a bidirectional order so that trends from the past as well as from the future can be learned. It also takes fewer learning parameters as compared to LSTM, which further takes less computational time. There are several steps involved in Gated-FCN. In the first step, spatial features and local trends are extracted. Noise is also removed using 8 layered FCN. Electricity load consumption time series is given as an input to FCN-8, unlike image processing where the 2D image is given as an input. Basically, the FCN is a variant of CNN. Therefore, all the mathematical formulation is derived from CNN [48]. The mathematical equation of convolutional layer is given as follows.

$$y_i := \sigma_i(w_i * x_i + b_i), \tag{6}$$

where y_i denotes the output of i^{th} convolutional layer. Similarly, σ represents the sigmoid function, w_i shows the weight factor, b_i is the bias factor and x represents the input of i^{th} convolutional layer. After the convolutional layers, the max pooling layers are integrated for the sake of dimensionality reduction. The mathematical equation of max pooling layer is given below.

$$y^m := max_{i,j \in R}(y_{i,j}). \tag{7}$$

In Equation 7, y^m shows the output of max pooling layer. i and j represent the i^{th} convolutional layer and j^{th} neuron, respectively. The finalized outcome of Equation 6 is further passed to Equation 7 for picking the potential features through max pooling operations. After applying the specified convolutional and pooling operations, the final result is devised in a 1D vector by using a flatten later. Finally, a fully connected layer is used to get the desired result. The mathematical equation of fully connected layer is given below.

$$y^f := g_i(w_i^f * y^m + b_i^f). \tag{8}$$

where y^f shows the final outcome of fully connected layer and g_i denotes the activation function. The mathematical formulation of EBiGRU [49] memory gates are given below.

$$r_t = \sigma(x_t * w_r + h_{t-1} * w_r + b_r). \tag{9}$$
$$u_t = \sigma(x_t * w_u + h_{t-1} * w_u + b_u). \tag{10}$$

where r_t and u_t denote the reset and update gates, respectively. x_t represents the input at time step t, w_r is the weight of reset gate, weight of the update gate is denoted by w_u. Whereas, b_r and b_u indicate the bias term of reset and update gate, respectively. The extracted local features and spatial correlation features are concatenated and fed to the enhanced version of EBiGRU, which learns long term temporal dependencies both from past and future. Moreover, it also represents the time series of electricity load consumption. L represents local features' trends of the time series data and I represents the input feature vector. LC_t represents concatenated local features and spatial correlation is represented as S_t. These concatenated features are used for separating temporal trends T_t based on long term dependencies. O_t represents a combined representation of spatial and temporal trends, as given in Equations 11 - 15, [49].

$$L_i = FCN(I), \tag{11}$$
$$LC_t = Concatenate(L_1, \ldots, L_i, \ldots, L_n), \tag{12}$$
$$S_t, T_t = EBiGRU(LC_t), \tag{13}$$
$$O_t = Concatenate(S_t, T_t). \tag{14}$$

In the second step, there are fusion layers for fusion of spatio-temporal features.

$$M_n = F((O_t)W_i, b_i), \quad i = 1, 2, \ldots, n. \tag{15}$$

M_n indicates the joint fusion of the learned spatial and temporal features. Whereas i represents i^{th} time window of input time series. To sum up the above-mentioned process, one dimensional FCN-8 extracts spatial correlation and local features. The concatenated spatial features are first used to capture temporal dependencies. Then, a fully connected dense layer is used for performing regression, which gives pointwise forecasting results [46]. Finally, the learned results are passed to the fully connected dense layer for predicting point values. The working of Gated-FCN is given in Algorithm 1. First of all, input features are corrupted by adding Gaussian noise in them. Then the corrupted features are input to FCN-8 for feature extraction. Afterward, the subsets of extracted features are created using Equation 16 in order to train snapshot models (demonstrated by the function BiGRU(.)) of EBiGRU. The prediction results are obtained using Equation 17. In the last, the average is taken for the results of all the snapshot models to obtain the final prediction results, as given in Equation 18. The simulation results are provided in Section IV.

$$(I_1', Y_1), \ldots, (I_N', Y_N) \tag{16}$$
$$Prediction_i = BiGRU_i(I_i', Y_i, XTest) \tag{17}$$
$$\text{EBiGRU-Prediction} = \frac{\sum_{i=1}^{n} Prediction_i}{N} \tag{18}$$

Algorithm 1 Working of Gated-FCN

1: **Begin**
2: **Inputs:** Input features X, Input targets Y, Test features $XTest$
3: **Output:** Predicted EC Y'
4: Add Gaussian noise to inputs $I = X + \sigma$
5: Extract features by FCN-8 as I'
6: Make N subsets of extracted features using Equation 16
7: **For** (i = 1 to N)
8: Make predictions using Equation 17
9: **End For**
10: Make final predictions using Equation 18
11: $Y' = EBiGRU\text{-}Prediction$
12: Return Y'
13: **End**

IV. SIMULATION RESULTS AND DISCUSSION

In this section, the simulation results and their discussion are given. The simulations are performed using the freely available Google Colab Repository. It allows easy access to write and execute Python codes using a web browser. It is a cloud server-based repository where the necessary hardware configurations are provided by Google cloud. Typically, 16 GB RAM and core i5 processor are available to perform experiments. Moreover, premier access is also available when more resources are required than the free version. In our scenario, all the experiments are conducted using the standard hardware resources. Moreover, a laptop having a core i5 processor with 8 GB RAM and 500 GB hard disk is used. All the code is written in Python language. Prior to the detailed discussion of simulation results, the dataset description is provided.

A. DATA DESCRIPTION

ISO-NE is an independent state organization that ensures region-wise electricity transmission. It ensures reliable transmission of energy in six states of North Eastern United States: Massachusetts, Maine, Connecticut, New Hampshire, Rhode Island and Vermont. Massachusetts is further divided into three load zones. Other states make their own load zone. In ISO-NE, every year approximately \$400 million transactions are performed by 7 billion customers. In order to verify the applicability of the proposed model, the ISO-NE dataset is used. It is a uni-variant dataset that comprises the load consumption data of 9 years, recorded from January 2011 to December 2019. The dataset is comprised of 77,760 samples [47]. It contains the dimensions of Date, Hour ending and consumed Load. In the date column, the date of EC is given, the Hour ending column demonstrates the specific hour of the day when energy is consumed and the consumed Load depicts the amount of load being consumed in an hour. The load is calculated in Megawatt per hour (MWh). The Date and Hour variables help in analyzing the patterns of consumed Load on weekdays and weekends. The analysis helps in efficient short-term load forecasting on the basis of users' past EC data. Using the data, simulations are performed.

B. VALIDATION SETTING OF THE PROPOSED SYSTEM MODEL

This section describes the validation setting of our proposed model. Table 2 describes the hyperparameters and layering structure of FCN model. Whereas, the validation setting of BiGRU is presented in Table 3.

TABLE 2. Validation setting of FCN model.

Hyperparameters	Values
No. of convolution layers	4
No. of pooling layers	4
No. of neurons in convolution layers	128, 128, 64, 64
Activation functions	ReLU, ELU, Leaky RELU, SELU
Dropout	0.2
Kernel size	3
No. of kernels	16
No. of dense layers	3
No. of fully connected layers	1

TABLE 3. Validation setting of BiGRU model.

Hyperparameters	Values
No. of hidden units	64
Activation functions	ReLU, ELU, Leaky RELU, SELU
Dropout	0.2
Batch size	128
Optimizer	Adam
Epochs	30

C. EXPLORATORY DATA ANALYSIS

In this study, the time horizons considered for forecasting and performing simulations are hourly, daily and weekly (short-term forecasting). Moreover, the presence of missing values is handled through the strong capabilities of convolutional and max-pooling layers of CNN. In convolving operation, the max-pooling layers pick up the maximum values and then discard the missing or zero values. Meanwhile, the abnormal values are tackled through BiGRU, which captures the long-term temporal correlation from users' load profiles. So, when an abnormal value appears, it is discarded by the reset gate of BiGRU. Furthermore, our model has the capability to learn non-linear complex relationships and temporal correlation from EC load data, which enable it to capture any uncertain change that occurs in the EC load profile. By having these properties, we claim that our model is consistent against any change that occurs in the older dataset. Fig. 6 represents load consumption of 9 years' hourly data from January 2011 to December 2019. Fig. 7 illustrates EC of weekdays and weekends. The results show that consumption on weekends is higher as compared to consumption on weekdays. Analysis of data is performed on daily basis in order to analyze the day-wise trends within the week. The consumption of a weekday: Monday, 7 January 2019, and a weekend: Sunday, 6 January 2019, is shown in Fig. 7. The graphical representation demonstrates that there exists a large difference between the EC of a weekend and a weekday from 04:00 to 16:00 hours.

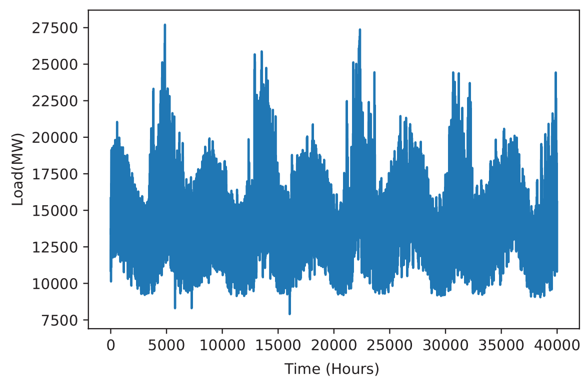


FIGURE 6. 2011-2019 ISO-NE hourly load consumption.

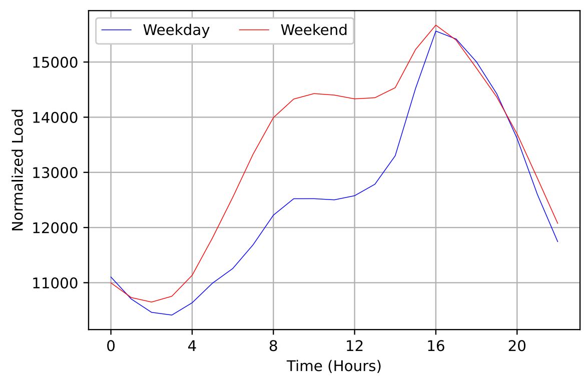


FIGURE 7. Weekend and weekday load consumption.

D. PREDICTIVE DATA ANALYSIS

The Gated-FCN deep learning model is validated by comparing its performance with different benchmark techniques. A detailed analysis of activation functions is performed in order to validate their impact on forecasting results. Multiple activation functions, such as ReLU, PReLU, ELU and SELU are used in the proposed model, as shown in Table 4.

TABLE 4. Exploring activation functions.

Activation Functions	Methods	RMSE (%)	MAPE (%)
RELU	Gated-FCN	4.87	5.02
	CNN-LSTM [14]	5.02	8.91
	BiGRU	7.01	12.03
	MLP	8.47	11.74
	LSTM	9.03	13.01
PRELU	Gated-FCN	4.03	5.11
	CNN-LSTM [14]	4.21	7.91
	BiGRU	6.83	10.73
	MLP	9.04	12.04
	LSTM	10.03	13.42
ELU	Gated-FCN	3.63	4.97
	CNN-LSTM [14]	4.93	8.03
	BiGRU	6.38	11.87
	MLP	8.14	12.51
	LSTM	10.53	14.12
SELU	Gated-FCN	2.87	3.01
	CNN-LSTM [14]	3.81	5.53
	BiGRU	4.97	8.35
	MLP	8.47	11.74
	LSTM	7.82	8.91

Among multiple activation functions, SELU performed the best for different state-of-the-art techniques. In order to validate Gated-FCN, two performance metrics are used: MAPE and RMSE. In the literature numerous Key Performance Indicators (KPIs) are available for performing accurate and reliable forecasting such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), etc. Among these, MAPE and RMSE are considered better than other KPIs. It is because RMSE emphasizes the most significant error whilst the MAE assigns equal importance to each error. Moreover, RMSE defines the square root of the average squared error. Hence, it is a better representative of the error as compared to MAE. In addition, MSE is easier to manipulate than RMSE but it is not representative of the original error as the error is squared. Whereas, the error is not squared in RMSE. In contrast, MAPE is easy to interpret and understand because it provides the error in terms of percentages. For instance, if MAPE is 5%, then it means the average difference between actual and forecasting value is 5%. That is the reason why MAPE is preferred over MAE. Hence, it is proved that RMSE and MAPE are better and sufficient measures for forecasting as compared to other traditional statistical measures. That is why most researchers use these two measures for evaluating the forecasting model [38]– [40]. The proposed model is compared with the state-of-the-art shallow machine learning and baseline deep learning models: FCN-8, LSTM, CNN-LSTM [14], BiGRU and Gated-FCN. The results demonstrate that the SELU activation function gives the best results when it is used in the Gated-FCN forecasting model with 2.87% RMSE and 3.01% MAPE. Whereas, ReLU activation function gives the worst accuracy with 4.87% RMSE and 5.02% MAPE. Hence, the SELU activation function is used for simulations. The simulation results are supported using both tabular and graphical representations. In this study, smaller values of MAPE and RMSE indicate better and accurate forecasting. It means that there is a high degree of precision in forecasting. In Fig. 8, it is observed that the proposed Gated-FCN model has the lowest MAPE for all days of the week as compared to existing models. The proposed model efficiently captures the hidden patterns and potential features through convolutional and pooling layers, respectively. Meanwhile, BiGRU learns the temporal correlations from the electricity load profile for efficient forecasting. Moreover, FCN-8 has high MAPE as compared to Gated-FCN, which means that a single model has weak forecasting precision. The performance of the proposed model over existing models is as follows. LSTM is vulnerable to overfitting issues as it becomes tedious to use the dropout mechanism to fix the issue. Moreover, the poor performance of LSTM happens when a number of memory cells are connected to the weights of the recurrent matrices. Also, LSTM is unable to learn patterns from arrays and memory indexing. CNN-LSTM has multiple layers that are time-consuming during operation. Also, CNN-LSTM loses its internal information as it sends the information to the neurons that are not fit to handle such information. Moreover, BiGRU also has a slow convergence issue, which needs to be tackled.

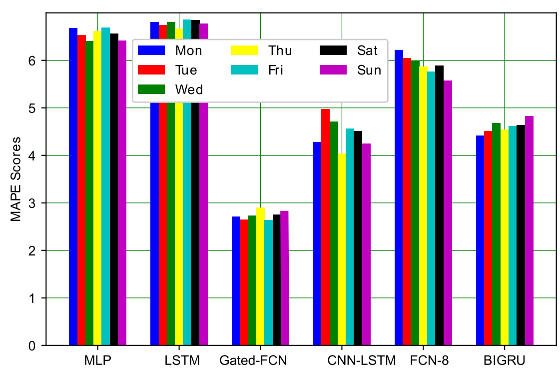


FIGURE 8. Load forecasting using MAPE.

Figs. 9 and 10 show forecasting results of 24 hours from the first Monday of January 2019 and the first week of January 2019. Results demonstrate the following trend of precision-accuracy: Gated-FCN > FCN-8 > CNN-LSTM > BiGRU > LSTM > MLP. It is analyzed that there are different trends of EC on different days of the week.

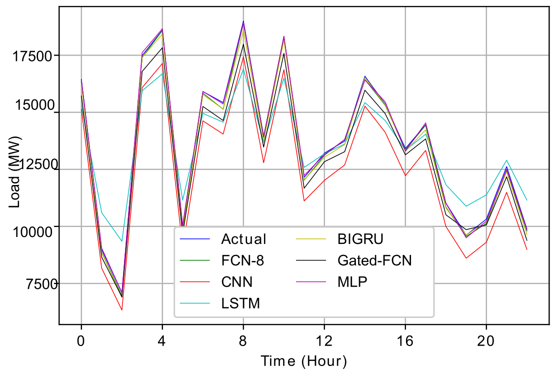


FIGURE 9. Daily load consumption forecasting.

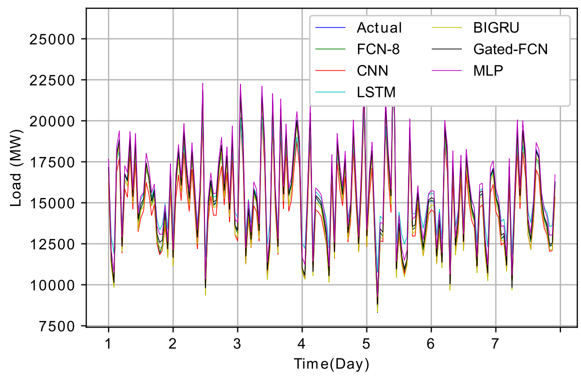


FIGURE 10. Weekly load consumption forecasting.

The tremendous feature learning abilities of CNN and the effective gated configuration of BiGRU make the proposed model superior in terms of efficient load forecasting as compared to other baseline models. In Fig. 9, the daily load forecasting is shown. It is seen from the figure that the actual data has the same pattern for both the proposed model and the existing models. It signifies that the proposed model not only memorizes the data but also learns the actual data. Unlike Gated-FCN, other existing models have problems of underfitting and overfitting. Moreover, the irregular patterns of data show the behavior of consumers during the winter season as they use heating appliances. It implies that load consumption is low at some hours of the day and vice versa.

Similar to Fig. 9, it is observed in Fig. 10 that other existing models are unable to learn the time series data as compared to the proposed Gated-FCN model. As the time series data changes from daily to weekly, the existing models become incapacitated to predict the data due to their inherent limitations. The same effect is seen in Figs. 11 - 17. The forecasting results of weekdays are shown in Figs. 11 - 15 and that of weekends are shown in Figs. 16 and 17. Moreover, from Fig. 17 it is observed that the proposed model has similar pattern with the actual data. It means that the model is able to address the vanishing gradient problem and preserve more historical information for accurate forecasting.

Fig. 18 and 19 show loss values of training and validation on the data. The results demonstrate the convergence progress

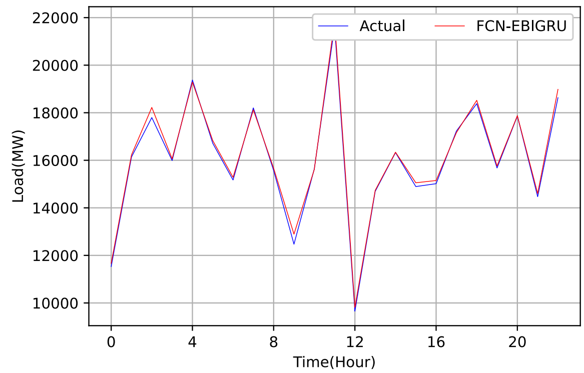


FIGURE 11. Load consumption - Monday.

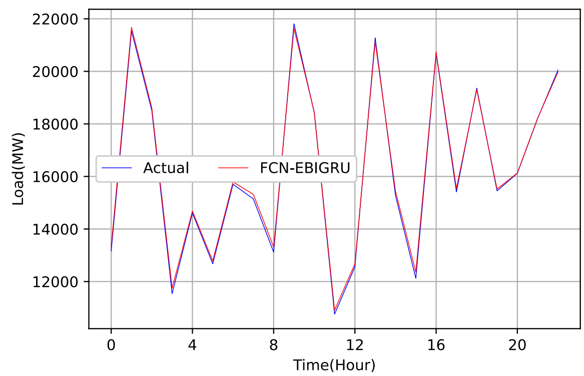


FIGURE 12. Load consumption - Tuesday.

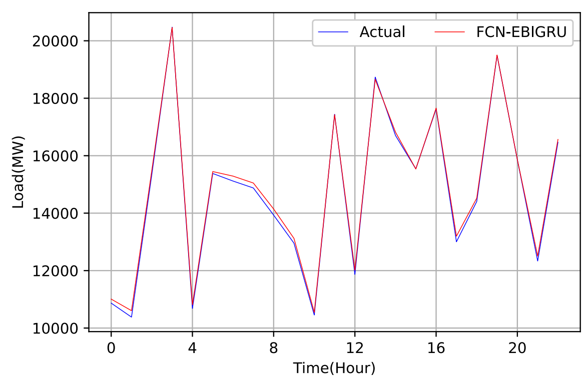


FIGURE 13. Load consumption - Wednesday.

of the Gated-FCN. It is observed that the benchmark forecasting technique, i.e., CNN-LSTM, has comparatively less convergence. In Fig. 18, the loss function of training and validation sets is presented. From the figure, it is observed that loss value diminishes as the number of epochs increases. It means that the number of variations of the actual and predicted values is negligible. Moreover, it is seen that there is a spike at the 49th epoch of the validation set. This is because of overfitting as the model tries to memorize the actual data. Nevertheless, the proposed model maintains its precision

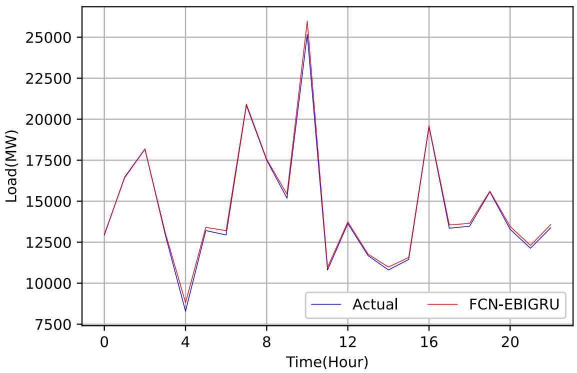


FIGURE 14. Load consumption - Thursday.

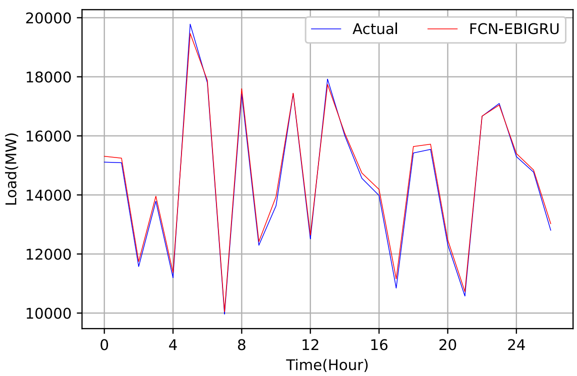


FIGURE 15. Load consumption - Friday.

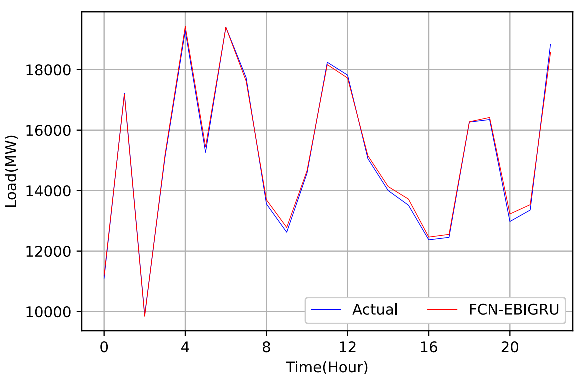


FIGURE 16. Load consumption - Saturday.

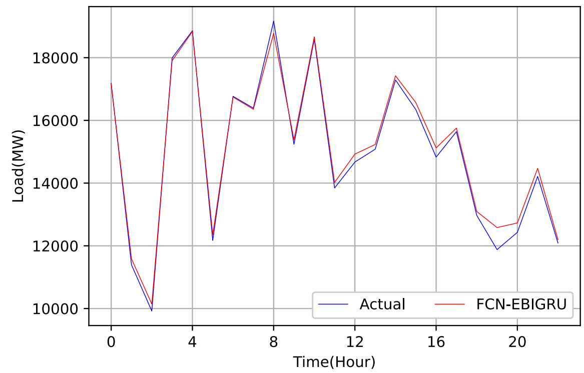


FIGURE 17. Load consumption - Sunday.

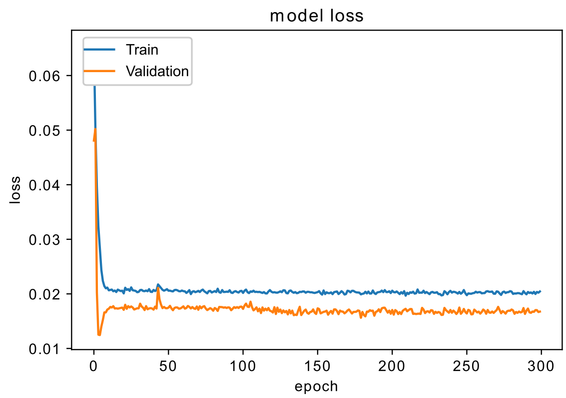


FIGURE 18. Loss of CNN-LSTM.

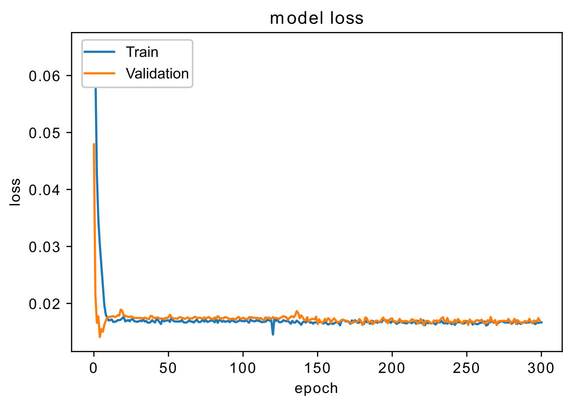


FIGURE 19. Loss of Gated-FCN.

during forecasting. Similar effect is shown in Fig. 19. However, at the 120th epoch, there is an issue of underfitting, which is resolved as the proposed model is further trained.

Tables 5-9 show forecasting errors from Monday to Friday. Whereas, Tables 10 and 11 show the forecasting errors for Saturday and Sunday, respectively. The RMSE of weekdays from Monday to Friday is 3.12, 3.37, 2.03, 3.14 and 2.82, respectively. Whereas, RMSEs for weekends, i.e., Saturday and Sunday are 3.96 and 2.04, respectively. The results show that Gated-FCN performed the best as compared to the other benchmark techniques because it learns both the global and the local characteristics optimally. The tabular results show that MAPE values of Gated-FCN and other state-of-the-art techniques: CNN-LSTM [14], FCN-8, LSTM, MLP and BiGRU. The results demonstrate that Gated-FCN gives the best precision accuracy a compared to other techniques. MLP, CNN and LSTM perform better than statistical techniques while forecasting EC load. To prove the nobility of the proposed forecasting technique, the statistical techniques, ARIMA, Simple Moving Average (SMA) and MLR, are

TABLE 5. Performance results - Monday.

Dataset	Day	Method types	RMSE (%)	MAPE (%)
ISO-NE	Monday	BiGRU	8.23	3.7
		CNN-LSTM [14]	5.92	3.63
		Gated-FCN	3.12	2.16
		FCN-8	9.15	4.63
		MLP	7.34	6.23
		LSTM	8.98	7.05
		ARIMA	14.33	11.21
		SMA	16.56	12.36
		MLR	11.36	9.81

TABLE 6. Performance results - Tuesday.

Dataset	Day	Method types	RMSE (%)	MAPE (%)
ISO-NE	Tuesday	BiGRU	7.74	4.95
		CNN-LSTM [14]	6.74	4.56
		Gated-FCN	3.37	2.52
		FCN-8	9.85	4.24
		MLP	8.36	6.72
		LSTM	8.78	6.93
		ARIMA	13.64	11.44
		SMA	15.23	12.34
		MLR	12.45	10.85

TABLE 7. Performance results - Wednesday.

Dataset	Day	Method types	RMSE (%)	MAPE (%)
ISO-NE	Wednesday	BiGRU	4.36	4.84
		CNN-LSTM [14]	3.26	4.74
		Gated-FCN	2.03	2.72
		FCN-8	9.12	4.72
		MLP	8.48	6.23
		LSTM	9.01	6.82
		ARIMA	13.78	11.67
		SMA	14.78	12.54
		MLR	12.14	11.43

used and compared with the proposed technique. However, metaheuristics and mixed linear integer programming based techniques are not considered because they cannot learn temporal correlation and complex relationship from long term EC load data. The above-mentioned statistical techniques

TABLE 8. Performance results - Thursday.

Dataset	Day	Method types	RMSE (%)	MAPE (%)
ISO-NE	Thursday	BiGRU	5.62	4.03
		CNN-LSTM [14]	4.85	4.12
		Gated-FCN	3.14	4.91
		FCN-8	9.97	5.25
		MLP	10.09	6.23
		LSTM	10.86	6.92
		ARIMA	12.31	11.53
		SMA	14.78	13.25
		MLR	13.59	12.69

TABLE 9. Performance results - Friday.

Dataset	Day	Method types	RMSE (%)	MAPE (%)
ISO-NE	Friday	BiGRU	4.13	5.92
		CNN-LSTM [14]	3.76	4.36
		Gated-FCN	2.82	2.46
		FCN-8	9.86	5.64
		MLP	8.91	6.24
		LSTM	9.01	6.92
		ARIMA	12.43	13.64
		SMA	14.34	11.45
		MLR	13.54	16.65

TABLE 10. Performance results - Saturday.

Dataset	Day	Method types	RMSE (%)	MAPE (%)
ISO-NE	Saturday	BiGRU	6.86	7.47
		CNN-LSTM [14]	5.48	7.06
		Gated-FCN	3.96	2.91
		FCN-8	10.69	5.36
		MLP	9.14	6.24
		LSTM	10.02	6.92
		ARIMA	14.54	12.09
		SMA	15.43	13.45
		MLR	12.43	11.36

TABLE 11. Performance results - Sunday.

Dataset	Seasons	Method types	RMSE(%)	MAPE(%)
ISO-NE	Sunday	BiGRU	7.04	12.63
		CNN-LSTM [14]	4.34	9.62
		Gated-FCN	2.04	5.61
		FCN-8	9.76	15.63
		MLP	8.32	16.12
		LSTM	9.52	14.34
		ARIMA	12.74	10.18
		SMA	13.65	11.54
		MLR	14.55	13.46

TABLE 12. Comparison of Gated-FCN with existing models using noise added consumption.

Models	Actual Consumption (RMSE (%))	Modified Consumption (RMSE(%))
BiGRU	8.23	8.52
CNN-LSTM [14]	5.92	6.14
Gated-FCN	3.12	3.29
FCN-8	9.15	9.25
MLP	7.34	7.68
LSTM	8.14	8.29
ARIMA	11.45	10.34
SMA	13.42	12.31
MLR	12.45	11.56

also perform poorly due to the long term dependencies. Moreover, these techniques do not have any memory module to memorize the essential patterns, trends and correlation of the electricity load profiles. As a result, these techniques produce inefficient EC forecast. From tabular results, it is seen that all of the statistical techniques have higher RMSE and MAPE as compared to other techniques.

Furthermore, in order to validate the robustness of Gated-FCN, the Gaussian noise is added to data with mean 0 and standard deviation 1. The results show that the proposed model outperforms existing models, as given in Table 12. The reason behind the efficient performance of Gated-FCN is the combination of the features of two deep learning models, which focuses on both temporal and spatial feature extraction and removal of the noise from the data. The results are also validated by graphical representation of loss function.

Tables 13 and 14 show the computational time and accuracy of Gated-FCN and other benchmark techniques, respectively. The computational time of Gated-FCN is relatively lower than other techniques, such as BiGRU, CNN-LSTM, LSTM, FCN-8, MLP, ARIMA, SMA and MLR. The statistical techniques have relatively high computational times because they are based on probability. In case of high dimensional data, they take longer time. Moreover, the existence of complex patterns in load profile slow down their performance while forecasting. Moreover, the accuracy of Gated -FCN is higher than other benchmark techniques.

TABLE 13. Computational time (sec).

Techniques	Training Time	Testing Time
Gated-FCN	192	78
BiGRU	156	65
CNN-LSTM [14]	67	45
LSTM	74	41
FCN-8	31	16
MLP	42	26
ARIMA	87	38
SMA	72	26
MLR	98	29

TABLE 14. Accuracy of forecasting techniques.

Techniques	Accuracy
BiGRU	95.6
CNN-LSTM [14]	96.14
Gated-FCN	98.6
FCN-8	95.087
MLP	95.36
LSTM	95.13
ARIMA	94.5
SMA	94.32
MLR	94.14

V. CONCLUSION AND FUTURE DIRECTIONS

This paper proposes a novel hybrid Gated-FCN model for efficient EC forecasting. It is a combination of two robust forecasting models: FCN-8 and EBiGRU. The former extracts spatial features from the data, identifies the variable that affects the EC forecasting and removes noise. Whereas, the latter extracts temporal features and predicts long-term temporal dependencies. It is also used to increase the network's robustness by averaging the weights of different training models. Moreover, the visual analysis of the data is performed, which shows different trends of EC on weekdays and weekends. The trends unveil consumers' consumption behavior patterns that help in producing electricity accordingly. To prove the efficacy of the proposed model, extensive simulations are performed. The proposed model is also compared with the existing models in terms of precision accuracy, MAPE and RMSE. The comparison results prove that Gated-FCN has the best precision accuracy and the least values of MAPE and RMSE as compared to FCN-8, BiGRU, CNN-LSTM, LSTM and MLP. Despite the proposed scheme being an ideal solution for efficient EC forecasting, it is computationally intensive because of using two deep learning models in a sequential manner. In the future, we will use different hyperparameter tuning techniques to enhance the performance of the proposed scheme in terms of efficient EC load forecasting. Furthermore, a multivariate data set will be considered to evaluate the importance of temperature in forecasting electricity demand and improve the forecasting accuracy.

.

AUTHORS' CONTRIBUTIONS

AQDAS NAZ is the first and main contributor of the article. She identified the problem statement and suggested comprehensive solutions to the identified problems. The entire implementation and simulations are performed by her. While NADEEM JAVAID is the corresponding author of the paper and the supervisor of AQDAS NAZ. The document is finalized under his keen supervision. In addition, MUHAMMAD SHAFIQ contributed by designing the proposed methodology while the mathematical formulation is done by JIN-GHOO-CHOI. The remaining authors worked on the representations of the simulation results. They have also discussed technical and logical reasoning of simulation results under the guidance of the principal author.

SUPPLEMENTARY MATERIAL

The data used in the proposed work is taken from ISO-NE, which is an independent state organization that deals with six states of North Eastern United States. The dataset can be retrieved from https://www.iso-ne.com/isoexpress/web/reports/load-and-demand.

REFERENCES

[1] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 841–851, Jan. 2019.

[2] M. Massaoudi, S. S. Refaat, I. Chihi, M. Trabelsi, F. S. Oueslati, and H. Abu-Rub, "A novel stacked generalization ensemble-based hybrid LGBM-XGB-MLP model for short-term load forecasting," *Energy*, vol. 214, Jan. 2021, Art. no. 118874.

[3] Y. Dong, X. Ma, and T. Fu, "Electrical load forecasting: A deep learning approach based on K-nearest neighbors," *Appl. Soft Comput.*, vol. 99, Feb. 2021, Art. no. 106900.

[4] S. M. J. Jalali, S. Ahmadian, A. Khosravi, M. Shafie-khah, S. Nahavandi, and J. P. S. Catalao, "A novel evolutionary-based deep convolutional neural network model for intelligent load forecasting," *IEEE Trans. Ind. Informat.*, vol. 17, no. 12, pp. 8243–8253, Dec. 2021.

[5] M. Ibrahim, A. Alsheikh, Q. Al-Hindawi, S. Al-Dahidi, and H. ElMoaqet, "Short-time wind speed forecast using artificial learning-based algorithms," *Comput. Intell. Neurosci.*, vol. 2020, pp. 1–15, Apr. 2020.

[6] K. Chen, K. Chen, Q. Wang, Z. He, J. Hu, and J. He, "Short-term load forecasting with deep residual networks," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3943–3952, Jul. 2019.

[7] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[8] G. Hafeez, K. S. Alimgeer, and I. Khan, "Electric load forecasting based on deep learning and optimized by heuristic algorithm in smart grid," *Appl. Energy*, vol. 269, Jul. 2020, Art. no. 114915, doi: 10.1016/j.apenergy.2020.114915.

[9] M. Alipour, J. Aghaei, M. Norouzi, T. Niknam, S. Hashemi, and M. Lehtonen, "A novel electrical net-load forecasting model based on deep neural networks and wavelet transform integration," *Energy*, vol. 205, Aug. 2020, Art. no. 118106, doi: 10.1016/j.energy.2020.118106.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," 2015, *arXiv:1506.04214*. [Online]. Available: http://arxiv.org/abs/1506.04214

[12] 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*, vol. 2, Aug. 2016, pp. 225–230.

[13] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," *IEEE Access*, vol. 6, pp. 1155–1166, 2018.

[14] T.-Y. Kim and S.-B. Cho, "Predicting residential energy consumption using CNN-LSTM neural networks," *Energy*, vol. 182, pp. 72–81, Sep. 2019.

[15] J. Chen, H. Jing, Y. Chang, and Q. Liu, "Gated recurrent unit based recurrent neural network for remaining useful life prediction of nonlinear deterioration process," *Rel. Eng. Syst. Saf.*, vol. 185, pp. 372–382, May 2019.

[16] W. Li, T. Logenthiran, and W. L. Woo, "Multi-GRU prediction system for electricity generation's planning and operation," *IET Gener., Transmiss. Distrib.*, vol. 13, no. 9, pp. 1630–1637, May 2019.

[17] M. S. Al-Musaylh, R. C. Deo, J. F. Adamowski, and Y. Li, "Short-term electricity demand forecasting with MARS, SVR and ARIMA models using aggregated demand data in Queensland, Australia," *Adv. Eng. Informat.*, vol. 35, pp. 1–16, Jan. 2018.

[18] G. Wang, Y. Su, and L. Shu, "One-day-ahead daily power forecasting of photovoltaic systems based on partial functional linear regression models," *Renew. Energy*, vol. 96, pp. 469–478, Oct. 2016.

[19] X. Zhang, J. Wang, and Y. Gao, "A hybrid short-term electricity price forecasting framework: Cuckoo search-based feature selection with singular spectrum analysis and SVM," *Energy Econ.*, vol. 81, pp. 899–913, Jun. 2019.

[20] M. S. AL-Musaylh, R. C. Deo, J. F. Adamowski, and Y. Li, "Short-term electricity demand forecasting using machine learning methods enriched with ground-based climate and ECMWF Reanalysis atmospheric predictors in southeast Queensland, Australia," *Renew. Sustain. Energy Rev.*, vol. 113, Oct. 2019, Art. no. 109293, doi: 10.1016/j.rser.2019.109293.

[21] L. Sehovac and K. Grolinger, "Deep learning for load forecasting: Sequence to sequence recurrent neural networks with attention," *IEEE Access*, vol. 8, pp. 36411–36426, 2020.

[22] J. Bedi and D. Toshniwal, "Deep learning framework to forecast electricity demand," *Appl. Energy*, vol. 238, pp. 1312–1326, Mar. 2019.

[23] K. Amasyali and N. M. El-Gohary, "A review of data-driven building energy consumption prediction studies," *Renew. Sustain. Energy Rev.*, vol. 81, pp. 1192–1205, Jan. 2018.

[24] Q. Dong, K. Xing, and H. Zhang, "Artificial neural network for assessment of energy consumption and cost for cross laminated timber office building in severe cold regions," *Sustainability*, vol. 10, no. 1, pp. 1–15, 2018.

[25] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*. [Online]. Available: http://arxiv.org/abs/1412.3555

[26] P. A. Mynhoff, E. Mocanu, and M. Gibescu, "Statistical learning versus deep learning: Performance comparison for building energy prediction methods," in *Proc. 8th IEEE PES Innov. Smart Grid Technol. Conf. Eur.*, Sep. 2018, pp. 1–6.

[27] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao, "Time series classification using multi-channels deep convolutional neural networks," in *Proc. Int. Conf. Web-Age Inf. Manage.* Macau, China: Springer, 2014, pp. 298–310.

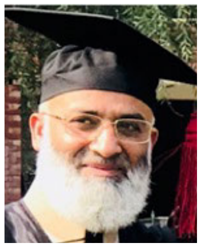
[28] H. Shi, M. Xu, and R. Li, "Deep learning for household load forecasting—A novel pooling deep RNN," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 5271–5280, Mar. 2017.

[29] W. Kong, Z. Y. Dong, D. J. Hill, F. Luo, and Y. Xu, "Short-term residential load forecasting based on resident behaviour learning," *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 1087–1088, Jan. 2018.

[30] S. Chang, Y. Zhang, W. Han, M. Yu, X. Guo, W. Tan, X. Cui, M. Witbrock, M. Hasegawa-Johnson, and T. S. Huang, "Dilated recurrent neural networks," 2017, *arXiv:1710.02224*. [Online]. Available: http://arxiv.org/abs/1710.02224

[31] J. Du, Y. Cheng, Q. Zhou, J. Zhang, X. Zhang, and G. Li, "Power load forecasting using BiLSTM-attention," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 440, no. 3, Feb. 2020, Art. no. 032115.
[32] Y. Wang, W. Liao, and Y. Chang, "Gated recurrent unit network-based short-term photovoltaic forecasting," *Energies*, vol. 11, no. 8, p. 2163, 2018, doi: 10.3390/en11082163.
[33] P. H. Kuo and C. J. Huang, "A high precision artificial neural networks model for short-term energy load forecasting," *Energies*, vol. 11, no. 1, p. 213, Jan. 2018, doi: 10.3390/en11010213.
[34] X. Ke, J. Zou, and Y. Niu, "End-to-end automatic image annotation based on deep CNN and multi-label data augmentation," *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 2093–2106, Aug. 2019.
[35] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomed. Signal Process. Control*, vol. 47, pp. 312–323, Feb. 2016.
[36] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4580–4584.
[37] N. Amral, C. S. Ozveren, and D. King, "Short term load forecasting using multiple linear regression," in *Proc. 42nd Int. Universities Power Eng. Conf.*, Sep. 2007, pp. 1192–1198.
[38] M. R. Haq and Z. Ni, "A new hybrid model for short-term electricity load forecasting," *IEEE Access*, vol. 7, pp. 125413–125423, 2019.
[39] S. Papadopoulos and I. Karakatsanis, "Short-term electricity load forecasting using time series and ensemble learning methods," in *Proc. IEEE Power Energy Conf. Illinois (PECI)*, Feb. 2015, pp. 1–6.
[40] D.-X. Niu, H.-F. Shi, and D. D. Wu, "Short-term load forecasting using Bayesian neural networks learned by hybrid Monte Carlo algorithm," *Appl. Soft Comput.*, vol. 12, no. 6, pp. 1822–1827, Jun. 2012.
[41] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, E. Howard, W. Hubbard, and L. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.
[42] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," 2015, *arXiv:1505.00853*. [Online]. Available: http://arxiv.org/abs/1505.00853
[43] T. Yang, Y. Wei, Z. Tu, H. Zeng, M. Kinsy, N. Zheng, and P. Ren, "Design space exploration of neural network activation function circuits," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 38, no. 10, pp. 1974–1978, Sep. 2018.
[44] D. She and M. Jia, "A BiGRU method for remaining useful life prediction of machinery," *Measurement*, vol. 167, Jan. 2021, Art. no. 108277, doi: 10.1016/j.measurement.2020.108277.
[45] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*, vol. 1. Cambridge, MA, USA: MIT Press, 2016.
[46] R. Zhao, D. Wang, R. Yan, K. Mao, F. Shen, and J. Wang, "Machine health monitoring using local feature-based gated recurrent unit networks," *IEEE Trans. Ind. Electron.*, vol. 65, no. 2, pp. 1539–1548, Feb. 2018.
[47] (2021). *ISO New England—Energy, Load, and Demand Reports*. [Online]. Available: https://www.iso-ne.com/isoexpress/web/reports/load-and-demand
[48] Z. Zheng, Y. Yang, X. Niu, H. N. Dai, and Y. Zhou, "Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids," *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1606–1615, Dec. 2017.
[49] R. Fu, Z. Zhang, and L. Li, "Using LSTM and GRU neural network methods for traffic flow prediction," in *Proc. 31st Youth Academic Annu. Conf. Chin. Assoc. Automat.*, Nov. 2016, pp. 324–328.

AQDAS NAZ received the bachelor's degree in software engineering from the University of Engineering and Technology, Taxila, Pakistan, and the master's degree in software engineering from the National University of Sciences and Technology, Pakistan. She is currently pursuing the Ph.D. degree with the Communications over Sensors (ComSens) Research Laboratory, COMSATS University Islamabad, Islamabad Campus, under the supervision of Dr. N. Javaid. Her research interests include data science, optimization, security and privacy, energy trading, and smart grid.



NADEEM JAVAID (Senior Member, IEEE) received the bachelor's degree in computer science from Gomal University, Dera Ismail Khan, Khyber Pakhtunkhwa, Pakistan, in 1995, the master's degree in electronics from Quaid-i-Azam University, Islamabad, Pakistan, in 1999, and the Ph.D. degree in computer science from the University of Paris-Est, France, in 2010. He is currently an Associate Professor and the Founding Director of the Communications over Sensors (ComSens) Research Laboratory, Department of Computer Science, COMSATS University Islamabad, Islamabad Campus. He has supervised 137 master's and 22 Ph.D. theses. He has authored over 900 articles in technical journals and international conferences. His research interests include energy optimization in smart grids and in wireless sensor networks using data analytics and blockchain. He was a recipient of the Best University Teacher Award from the Higher Education Commission of Pakistan, in 2016, and the Research Productivity Award from Pakistan Council for Science and Technology, in 2017. He is an Associate Editor of IEEE Access and an Editor of the *Sustainable Cities and Society* journal.

MUHAMMAD ASIF received the B.S. degree in information technology from the University of Gujrat, Gujrat, Pakistan, in 2017. He is currently pursuing the M.S. degree in computer science with the Communications over Sensors (ComSens) Research Laboratory, Department of Computer Science, COMSATS University Islamabad, Islamabad, Pakistan, under the supervision of Dr. N. Javaid. His research interest includes electricity theft detection using machine learning techniques.

MUHAMMAD UMAR JAVED received the bachelor's and master's degrees in electrical engineering from the Government College University Lahore, Lahore, Pakistan, in 2014 and 2018, respectively. He is currently pursuing the Ph.D. degree with the Communications over Sensors (ComSens) Research Laboratory, Department of Computer Science, COMSATS University Islamabad, Islamabad, under the supervision of Dr. N. Javaid. He has authored more than 12 research publications in international journals and conferences. His research interests include smart grid, electric vehicles, and blockchain.

ABRAR AHMED received the B.S. degree in computer engineering from COMSATS University Islamabad, (Formerly, COMSATS Institute of Information Technology), Pakistan, in 2006, the M.S. degree from Lancaster University, U.K., in 2008, and the Ph.D. degree in electrical engineering from COMSATS University Islamabad, in 2017. Since 2006, he has been associated with COMSATS University Islamabad, where he currently holds the position of an Assistant Professor. His research interests include wireless channel modeling and characterizing, smart antenna systems, nonorthogonal multiple access techniques, and adaptive signal processing.

SARDAR MUHAMMAD GULFAM received the master's degree from Tampere University of Technology, Finland, in 2010, and the Ph.D. degree from COMSATS University Islamabad, Islamabad, Pakistan, in 2017. He is currently working as an Assistant Professor with the Department of Electric and Computer Engineering, COMSATS University Islamabad.

MUHAMMAD SHAFIQ received the master's degree in information technology (IT) from the University of the Punjab, Gujranwala, Pakistan, in 2006, the M.S. degree in computer science from the University Institute of Information Technology, Arid Agriculture University, Rawalpindi, Pakistan, in 2010, and the Ph.D. degree in information and communication engineering from Yeungnam University, South Korea, in February 2018. He was with the Faculty of Computing and IT, University of Gujrat, Gujrat, Pakistan, as a Faculty Member, from 2010 to 2014, and formerly held the same position with the Department of Computer Science and IT, Federal Urdu University, Islamabad, Pakistan. His research interests include the Internet of Things (IoT), cognitive radio-based IoT networks-architecture and design, mobile *ad-hoc* networks, wireless sensor networks, performance, management, and security, 5G cellular networks, admission control, mobility management, device-to-device communications, medium access control protocols, internet routing protocols, spectrum trading and auctions, information systems, design, and access control, and human–computer interaction.

JIN-GHOO CHOI received the Ph.D. degree from the School of Electrical Engineering and Computer Science, Seoul National University, in 2005. From 2006 to 2007, he was a Senior Engineer with Samsung Electronics. In 2009, he was with the Department of Electrical and Computer Engineering, The Ohio State University, as a Visiting Scholar. In 2010, he joined the Department of Information and Communication Engineering, Yeungnam University, as a Faculty Member. His research interests include performance analysis of communication networks, resource management in wireless networks, and wireless sensor networks.