IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 8, 2020, accepted May 26, 2020, date of publication June 15, 2020, date of current version June 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3002380

ECMCRR-MPDNL for Cellular Network Traffic Prediction With Big Data

VENKATA SUBBARAJU DOMMARAJU[1], (Member, IEEE), KARTHIK NATHANI[1], (Member, IEEE), USMAN TARIQ[2], FADI AL-TURJMAN[3], SURESH KALLAM[4], (Member, IEEE), PRAVEEN KUMAR REDDY M[5], AND RIZWAN PATAN[6]

[1]Department of Computer Science, University of the Cumberlands, Williamsburg, KY 40769, USA
[2]College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia
[3]Research Centre for AI and IoT, Artificial Intelligence Engineering Department, Near East University, 99138 Nicosia, Turkey
[4]Department of Computer Science and Engineering, Sree Vidyanikethan Engineering College, Tirupati 517102, India
[5]School of Information Technology and Engineering, Vellore Institute of Technology, Vellore 632014, India
[6]Department of Computer Science and Engineering, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada 520007, India

Corresponding author: Usman Tariq (u.tariq@psau.edu.sa)

ABSTRACT Big data comprises a large volume of data (i.e., structured and unstructured) stored on a daily basis. Processing such volume of data is a complex task as well as the challenging one. This big data is applied in the cellular network for traffic prediction. Now, benefiting from the big data in cellular networks, it becomes possible to make the analyses one step further into the application level. In order to improve the traffic prediction accuracy with minimum time, Expected Conditional Maximization Clustering and Ruzicka Regression-based Multilayer Perceptron Deep Neural Learning (ECMCRR-MPDNL) technique is introduced. The ECMCRR-MPDNL technique initially collects a large volume of data over the spatial and temporal aspects of cellular networks. Then the collected data are trained with multiple layers such as one input layer, two hidden layers, and one output layer. The activation function is used at the output layer to predict the network traffic based on the similarity value with higher accuracy. These predictors are evaluated using real network traces. Finally, the error rate is calculated for minimizing the prediction error. Experimental evaluation is carried out using a big dataset with different metrics such as prediction accuracy, false-positive and prediction time. The observed result confirms that the proposed ECMCRR-MPDNL technique improves on an average the 98% of performance of network traffic prediction with higher accuracy and 20 % minimum time as well as the false-positive rate as compared to the state-of-the-art methods.

INDEX TERMS Big data, cellular network traffic prediction, multilayer perceptron deep neural learning, iterative expected conditional maximization clustering, Ruzicka similarity, regression, activation function.

I. INTRODUCTION

With the increasing trend of mobile operators and internet access, the data traffic has posed great challenges since the load of the network is constantly increased. Therefore, the network traffic analysis and prediction are an essential part for cellular networks to minimize the load, since it is used for network control and management as well as service provisioning. The several works have been designed in the cellular network traffic prediction, but it has some challenges due to the large volume of temporal and spatial dynamics introduced by different user Internet behaviors.

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

A Spatial-Temporal Cross-domain neural Network (STCNet) was developed in [1] to increase the cellular network traffic prediction with complex patterns. The designed model uses the clustering concept to divide the city into different zones. Though the model reduces the mean square error, the performance of traffic prediction accuracy was not calculated. A Graph Neural Network with Decomposed Cellular Traffic model (GNN-D) was developed in [2] to improve the cellular traffic prediction by learning the spatial and temporal dependencies. The designed model failed to minimize the prediction time of continuously evolving traffic patterns.

Extending Labeled Data (ELD) was introduced in [3] to discover the label of unknown mobile network traffic.

The model failed to minimize the mobile network traffic prediction error. A traffic pattern extraction and modeling method were introduced in [4] for processing big cellular data. The method uses the clustering concept to minimize the complexity of prediction, but the accurate traffic prediction was not performed. A Jordan recurrent neural network (JNN) using a firefly algorithm was introduced in [5] to forecast the cellular data traffic with minimum error. The time complexity of traffic prediction was not minimized.

A three-layer classifier using machine learning was developed in [6] to discover mobile traffic with higher precision. Though the method reduces the false positive rate, the prediction time was not minimized. An application-level traffic prediction method was introduced in [7] using traffic big data. The designed method failed to enhance traffic prediction accuracy.

An Exponential Smoothing Method was developed in [8] to predict the cellular network traffic with lesser complexity. But the error rate was not reduced. A multiple RNN based learning models were developed in [9] using a unified multi-task learning method for enhancing the traffic forecasting by using Spatio-temporal correlations among base stations. Though the method minimizes the error rate, the accuracy was not improved. A cellular traffic offloading problem was resolved in [10] by the link prediction with minimum delay. But it failed to analyze the multiple data effectively from the cellular network.

A. CONTRIBUTIONS

To solve the issues identified from the above-said literature, a novel technique called the ECMCRR-MPDNL technique is introduced. The major contribution of the paper is summarized as follows,

- To improve the cellular network traffic prediction accuracy, an ECMCRR-MPDNL technique is introduced by learning the given data using different layers. In the hidden layer, the data within the clusters are analyzed using the Ruzicka regression function. The regression measures the similarity between the data and traffic patterns. Then the similarity results are transferred into the output layer and analyzing the results with the threshold value using activation function. If the similarity value is higher than the threshold, then the traffic density is correctly predicted at the particular base station.
- To minimize the traffic prediction error rate i.e. false positive rate, Multilayer Precepted Deep Neural Learning effectively providing accurate results with a lesser mean square error. Based on the Ruzicka similarity values, the higher possibility of traffic is correctly predicted at the output layer.
- To minimize the cellular network traffic prediction time, the ECMCRR-MPDNL technique uses the Iterative Expectation conditional maximization clustering technique. The clustering technique is used for partitioning the total network data into different groups based on the condition likelihood probability between the spatial data. With the clustering results, the traffic prediction is carried out using the regression analysis. This helps to minimize the time complexity of cellular network traffic prediction.

B. MOTIVATION

The network traffic analysis and prediction are an essential part for cellular networks to minimize the load, since it is used for network control and management as well as service provisioning. Afterwards, with the aid of the traffic ''big data'', To design a model to prediction framework of cellular network traffic. To achieve prediction accuracy, false-positive and prediction time.

C. OUTLINE OF PAPER

The structure of the paper is organized into different sections. Section 2 discusses the related works of traffic data prediction. In section 3, a description of the proposed ECMCRR-MPDNL technique is presented with a neat diagram. Section 4 presents an experimental scenario with a big cellular traffic dataset. Section 5 evaluates the performance results of the proposed techniques and existing algorithms in terms of traffic prediction accuracy, false positive rate and traffic prediction time. Section 6 concludes the paper.

II. RELATED WORK

A multivariate Long Short-Term Memory (LSTM) algorithm was designed in [11] to predict the traffic networks by performing the call detail record (CDR) data analysis. The designed algorithm failed to consider the large volume of wireless data and complex data types for network prediction.

A Deep Belief Network was developed in [12] to forecast the network traffic using spatiotemporal correlation with minimum error rate. But the traffic prediction time complexity was not minimized. A Deep Traffic Predictor (DeepTP) was developed in [13] to forecast the traffic from the cellular network. The model consumed more time for traffic prediction.

A deep-learning-based Cloud Radio Access Network (C-RAN) optimization technique was developed in [14] using the Multivariate LSTM model to perform traffic forecasts based on temporal dependence and spatial correlation. But the performance of the deep learning was not improved by considering different traffic patterns. A clustering-based artificial neural network (C-ANN) model was introduced in [15] for classifying the mobile traffic patterns. Though the model increases the accuracy, the false positive rate was not minimized.

A novel approach was introduced in [16] that initiate the users to smooth the traffic temporally. But the method failed to ensure higher prediction accuracy. A measurement-driven model was developed in [17] for mobile data traffic prediction using big data collected from a crowded metropolitan area. The traffic parameters spatial correlation was not analyzed to further enhance traffic prediction.

A densely connected Convolutional Neural Networks (CNN) was introduced in [18] to forecast the cellular network traffic with minimum error. But it failed to achieve better performance using a large amount of training data.

A Group Method of Data Handling (GMDH) polynomial neural network was developed in [19] to accurate traffic prediction. However, the time complexity of the traffic prediction was not minimized. Various machine learning and statistical methods were developed in [20] for forecasting the voice traffic of the Mobile Telecommunication System. However, the performance of the accurate prediction was not obtained with minimum complexity.

Many real network traces are evaluated for predicting. Evaluation of accuracy and cost, both in terms of power consumption and computation complexity, is offered. There is an observation over a double exponential smoothing predictor provides a reasonable tradeoff between cost overhead and performance [21]. However, it is fails to address the prediction accuracy, false-positive and prediction time.

In [22]–[25] techniques are used traffic flow prediction for vehicles networks and network applications. However, they are used pattern for predicting traffic using deep learning architectures can capture these nonlinear spatio-temporal effects. Methodology on real sensor data for predict traffic flows during two special events [26]; a Chicago Bears football game and an extreme snowstorm event. Both situations have shrill traffic flow regime changes, happening very speedily, and how deep learning offers [27] accurate short-term traffic flow predictions. However, it fails to address the traffic in different scenarios and even it fails prediction accuracy [28], false-positive and prediction time.

The above-said issues and challenges are overcome by introducing an ECMCRR-MPDNL technique. The description of the ECMCRR-MPDNL technique is presented in the following section.

III. METHODOLOGY

With billions of mobile devices accessing the Internet, cellular traffic has grown extremely in the past few years. Large volumes of data about the cellular traffic are collected at cell towers (i.e. base station) that widely implemented for daily network management. With the increasing volume of big cellular data, traffic prediction plays a challenging one due to temporal and spatial dynamics established by different user behavior.

The prediction is the statistical technique used for extracting more relevant information from the large volume of data and predicting future outcomes by collecting and analyzing the current and past events [29]. Accurate cellular network traffic prediction at base station enables to ensure good quality of services. Based on this motivation, cellular network traffic prediction is carried out by employing the deep learning and statistical learning concept. Therefore, the ECMCRR-MPDNL technique is introduced to handle the large volume of cellular data for accurate prediction by deeply learning the higher-level features from the raw-datasets using multiple layers [30]. The flow process of the ECMCRR-MPDNL technique is shown in figure 1.

Figure 1 illustrates the flow process of the proposed ECMCRR-MPDNL technique to obtain better traffic

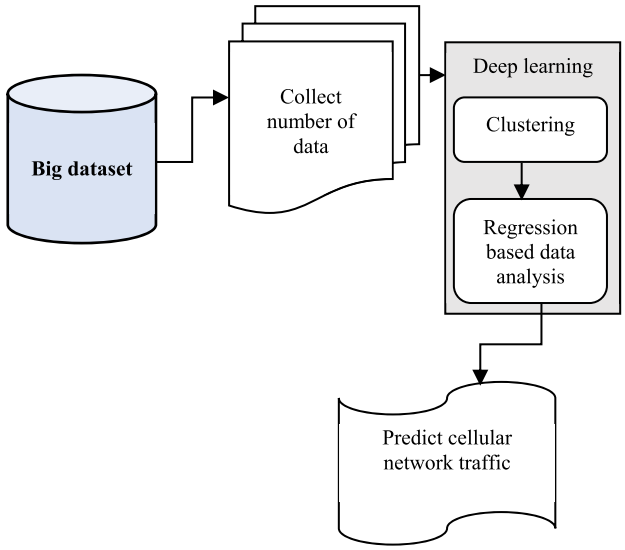

FIGURE 1. Flow process of ECMCRR-MPDNL technique.

prediction with minimum time. Initially, the big dataset is taken as input for predictive analytics. The numbers of spatial and temporal data are collected from the big dataset. After that, the input data analyzed using feedforward multilayer percepted deep neural learning. Then the input data is transferred in only one direction from the input nodes of the deep architecture and then analyzed in the hidden layers using regression function. Finally, the prediction results are obtained at the output layer. The structure of the feedforward multilayer percepted deep neural learning technique is shown in figure 2.

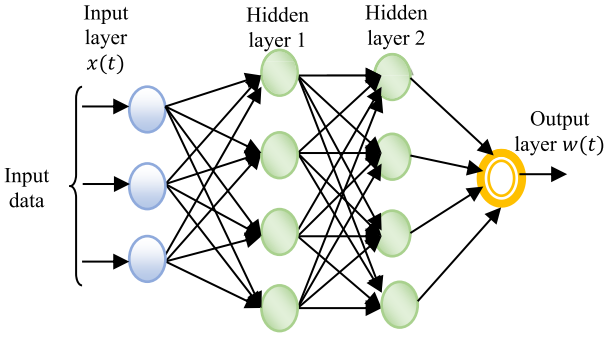

FIGURE 2. Diagram of the feedforward multilayer percepted deep neural learning.

Figure 2 depicts the structure of the feed-forward multilayer percepted deep neural learning with the one input layer, two hidden layers, and one output layer. The spatial and temporal data are given to the input layer d_{st_1}, d_{st_2}, d_{st_2}, ...d_{st_n} at d_{st_n} is d indicated data, s indicates source, t indicated time and n indicates number of instances. The feed-forward multilayer percepted deep neural learning comprises the neurons-like the nodes are fully connected to the consequent layers with the adjustable weights and performs the deep learning of

input data. Then the inputs are transformed into first hidden layers. The adjustable weight between the input and hidden layer is represented as ρ_1.

$$x(t) = \sum_{i=1}^{n} \mathrm{d}_{\mathrm{st_i}} * \rho_1 \tag{1}$$

where, $x(t)$ denotes an input with time 't', $\mathrm{d}_{\mathrm{st_i}}$ represents a big data, ρ_1 denotes an adjustable weight between the input and first hidden layer. In the first hidden layer, the clustering process is carried out to divide the network data into dissimilar groups for accurately predicting the traffic patterns in the given location with minimum time. The iterative expected conditional maximization clustering technique is applied for clustering the total network data into different groups. In the second hidden layer, Ruzicka regression is applied for analyzing the input data within the clusters. The regression is a statistical measurement used to find the strength of the relationship between the variables (i.e. data and traffic patterns). The proposed clustering technique is a statistical model to compute the Likelihood between the data. The likelihood expresses how one data is more like others. The iterative method is a process to provide efficient solutions for a class of particular problems while handling a large number of data.

The input data $d_{st_1}, d_{st_2}, d_{st_2}, \ldots . d_{st_n}$ are taken from the dataset and transferred from the input layer to the first hidden layer for performing the clustering process. By applying the Iterative Expectation conditional maximization clustering, 'm' number of clusters $s_1, s_2, s_3, \ldots, s_m$ and cluster centers $\alpha_1, \alpha_{2,}\alpha_{3,\ldots\ldots}\alpha_m$ are initialized randomly. Then the expected likelihood between the input data and cluster center is computed for grouping the data into clusters.

$$e_p\left(\mathrm{d}_{\mathrm{st_i}}, \alpha_j\right) = \sum \log \frac{p_r\left(d_{st_i} \mid p_r(\alpha_j)\right) * p_r\left(\alpha_j\right)}{p_r\left(d_{st_i}\right)} \quad j = 1, 2, 3, \ldots, m, i = 1, 2, 3, \ldots, n \tag{2}$$

where, $\mathrm{e_p}\left(\mathrm{d}_{\mathrm{st_i}}, \alpha_{\mathrm{j}}\right)$ denotes an expected likelihood between the data '$\mathrm{d}_{\mathrm{st_i}}$' and cluster center α_j, p_r denotes a probability that is used to find similar data to the cluster center. The likelihood is maximized which is expressed as follows,

$$M_L = \arg\max e_p\left(\mathrm{d}_{\mathrm{st_i}}, \alpha_j\right) \tag{3}$$

$$M_L = \arg\max \sum \log \frac{p_r\left(d_{st_i} \mid p_r(\alpha_j)\right) * p_r(\alpha_j)}{p_r(d_{st_i})} \tag{4}$$

where, M_L denotes maximum likelihood estimators which maximize the likelihood value determined from the expectation step. An arg *max* denotes an abbreviation of the argument of the maximum function. Therefore, the condition likelihood probability indicates that the more similar data are grouped into the particular cluster. This process is iterated until all the data is grouped into different clusters. As a result, the clustering process minimizes the traffic prediction time in the cellular network. The clustering output is fed into the hidden layer 2.

In the hidden layer 2, the Ruzicka regression is applied for analyzing the input data within the clusters. The regression is a statistical measurement used to find the strength of the relationship between the variables (i.e. data and traffic patterns). The relationship is measured using Ruzhika similarity coefficient, which is mathematically expressed as follows,

$$\gamma = \frac{c_d \cap g}{\sum c_d + \sum g - c_d \cap g} \tag{5}$$

where, γ represents a Ruzicka similarity coefficient, c_d represents data within the cluster g denotes traffic patterns, $c_d \cap g$ denotes a mutual dependence between the two variables. Here the traffic patterns are defined by the number of transferred data connected with a particular base station in a specific time. The similarity coefficient (γ) provides a value between 0 and 1 [0, 1]. Therefore, the output of the hidden layer is expressed as follows,

$$z(t) = x(t) + \rho_h + b(t-1) \tag{6}$$

where, $z(t)$ represents a hidden layer output at a time 't', 'b' denotes output from the hidden layer and '(t−1)' indicates previous layer, ρ_h represents a weight between the two hidden layers, $x(t)$ represents the input. The output of the second hidden layer is given to the output layer. In the output layer, the similarity values are analyzed for predicting the network traffic with minimum time.

$$w(t) = \delta\left(\rho_2 * z(t)\right) \tag{7}$$

where, $w(t)$ represents the output of the deep neural learning, ρ_2 denotes an adjustable weight between the hidden layer and output layer, $z(t)$ denotes an output of the hidden layer, δ denotes an activation function. The activation function is used for predicting the network traffic based on the similarity value.

$$\delta = \begin{cases} \gamma > \vartheta & \textit{higher possibility for traffic occurances} \\ \gamma < \vartheta & \textit{Otherwise} \end{cases} \tag{8}$$

where δ denotes an activation function, ϑ denotes a threshold value of the Ruzicka similarity coefficient. If the similarity value is higher than the threshold value, there is a higher possibility of traffic occurrences. In this way, the network traffic is accurately predicted in a given location. After that, the mean square error rate is computed based on the squared difference between the actual and the observed results. It is mathematically expressed as follows,

$$Error = \{w_a(t) - w(t)\}^2 \tag{9}$$

where, $w_a(t)$ represents an actual prediction output, $w(t)$ represents the observed traffic prediction results using the proposed technique. Based on the error value, the weights between the layers are adjusted and this process is repeated until it finds the minimum error.

$$argmin = \arg\min(Error) \tag{10}$$

where arg *min* denotes an argument of the minimum function to find the minimum error of the observed results. This process minimizes the incorrect traffic prediction in a cellular

Algorithm 1 Expected Conditional Maximization Clustering and Ruzhika Regression-Based Multilayer Precepted Deep Neural Learning

Input: Big dataset, Number of data $d_{st_1}, d_{st_2}, d_{st_2}, \ldots . d_{st_n}$.
Output: Improve traffic prediction accuracy in
Begin

1. Number of d_{st_i} taken as input at a time 't' i.e. $x(t)$ ——–**input layer**
2. Partition d_{st_i} data into different clusters $s_1, s_2, s_3, \ldots, s_m$\\ **Hidden layer** 1
3. Initialize number of clusters $s_1, s_2, s_3, \ldots, s_m$ and centers $\alpha_1, \alpha_{2,}\alpha_{3,\ldots\ldots}\alpha_m$
4. **for each** d_{st_i}
5. **for each** α_m
6. Compute expected probability $e_p\left(d_{st_i}, \alpha_j\right)$
7. Maximize expected probability for grouping similar data to the particular cluster
8. **end for**
9. **end for**
10. **For each** data c_d within the cluster
11. Compute the similarity between c_d and g \\ **Hidden layer** 2
12. **End for**
13. Transfer $z(t)$ into $w(t)$
14. **If** $(\gamma > \vartheta)$ **then** **output layer**
15. Higher possibilities of traffic occurrences
16. **Else**
17. No traffic
18. **End if**
19. Compute the error E
20. Update the weights between the layers
21. **The p**rocess is repeated until find minimum error
22. **End for**
23. **End**

network with higher accuracy. The algorithmic process of the is described as follows,

Algorithm 1 describes the step by step process of cellular network traffic prediction using feedforward multilayer percepted deep neural learning with higher accuracy. The big dataset is considered as input to the deep neural learning. The input data are divided into different groups based on the location information by applying the Iterative Expectation conditional maximization clustering technique. After clustering, the regression analyzes is carried out between the data within the cluster and the traffic patterns at the second hidden layer using Ruzhika similarity measure. At the output layer, the similarity values are verified with the threshold value. The similarity value is greater than the threshold, and then the technique correctly predicts the possibility of traffic occurrences at a particular location. Finally, the mean square error is calculated for the observed results and the process is continued until the minimum error is attained. As a result, deep neural learning technique improves the network traffic prediction accuracy and minimizes the false positive rate.

IV. EXPERIMENTAL SETUP AND PARAMETER SETTINGS

An experimental evaluation of the proposed ECMCRR-MPDNL technique and existing methods STCNet [1], GNN-D [2] are carried out using Java Language with City Cellular Traffic Map (C2TM) dataset taken from the https://github.com/caesar0301/city-cellular-traffic-map. The C2TM dataset comprises the two files such as traffic and topology. The traffic file contains the 1625680 rows and 5 columns (i.e. attributes). The topology file contains the 13296 rows and 3 columns.

TABLE 1. Attributes information for traffic trace file.

Attribute name	Description
BS	Identity of each cellular base station in the public data
Time_hour	Hourly timestamp in UNIX epoch time (time zone GMT+8).
Users	Number of active users associated with specific base station and hour
Packets	Number of transferred packets associated with specific base station and hour
Bytes	A number of transferred bytes associated with a specific base station and hour.

TABLE 2. Attributes information for topology file.

Attribute name	Description
BS	Identity of each cellular base station in this public data
Lon	Relative longitude of the given base station
Lat	Relative latitude of the given base station

In table 1 and 2, the time dimension is represented by the Gregorian calendar, with the starting point of each hour to denote the following 60 minutes. The relative location of the base station is in longitude/latitude to provide some analysis. The packets column in the traffic trace file is used for discovering the traffic density over space. The experimental evaluation is done with the above said datasets for predicting the network traffic with different parameters are listed below,

- Traffic prediction accuracy
- False-positive rate
- Prediction time.

V. EXPERIMENTAL SETUP AND PARAMETER SETTINGS

The results analysis of the proposed ECMCRR-MPDNL technique is compared to existing methods STCNet [1], GNN-D [2] with different performance metrics such as Traffic prediction accuracy, False-positive rate and prediction time. The description of various is presented in each section and the comparative results are obtained using table and graphical representation.

A. TRAFFIC PREDICTION ACCURACY

Traffic prediction accuracy is referred to as a ratio of a number of data correctly predicted as a possibility of traffic occurrences or not to the total number of data taken as input

for conducting the experiment. The formula for calculating the traffic prediction accuracy is calculated as follows,

$$TPR = \left(\frac{Number\ of\ data\ correctly\ predicted}{n}\right) * 100 \quad (11)$$

where n denotes the number of data. The prediction accuracy is measured in terms of percentage (%).

Sample calculation:

- **Proposed ECMCRR-MPDNL:** Number of data taken as input is 500 and the number of data correctly predicted as traffic occurrences or not is 450. Then the traffic prediction accuracy is mathematically calculated as follows,

$$TPR = \left(\frac{450}{500}\right) * 100 = 90\%$$

- **Existing STCNet:** Number of data taken as input is 500 and the number of data correctly predicted as traffic occurrences or not is 430. Then the traffic prediction accuracy is mathematically calculated as follows,

$$TPR = \left(\frac{430}{500}\right) * 100 = 86\%$$

- **Existing GNN-D:** Number of data taken as input is 500 and the number of data correctly predicted as traffic occurrences or not is 410. Then the traffic prediction accuracy is mathematically calculated as follows,

$$TPR = \left(\frac{410}{500}\right) * 100 = 82\%$$

Table 3 shows the performance results of traffic prediction accuracy using three different techniques namely ECMCRR-MPDNL technique, STCNet [1] and GNN-D [2]. For calculating the traffic prediction accuracy, the numbers of spatial and temporal data are collected from the C2TM dataset. The number of data has been taken in the range from 500 to 5000 for calculating the traffic prediction accuracy. While handling the large volume of data, the accurate traffic prediction is higher using the ECMCRR-MPDNL technique than the other two techniques is shown in table 3. The graphical representation of the traffic prediction accuracy with different inputs is shown in figure 3.

The graphical results of traffic prediction accuracy with various cellular data are shown in figure 3. In the graph, the number of data as taken as input in the 'x' axis whereas the prediction accuracy is obtained at the 'y' axis. The prediction accuracy of three different techniques ECMCRR-MPDNL technique, STCNet [1] and GNN-D [2] represented by three different colors namely blue, red and green respectively. From the figure, it is inferred that the traffic prediction accuracy is increased using the ECMCRR-MPDNL technique as compared to the other two methods. The reason behinds the ECMCRR-MPDNL technique effectively performs the regression analysis at the hidden layer for analyzing the data with the traffic patterns. The regression function measures the similarity of data and traffic patterns in order to identify the traffic density at a particular location. The activation function correctly predicts the possibility of traffic occurrences with the help of similarity value at a particular space. As a result, the regression analysis accurately predicts cellular network traffic with big data.

TABLE 3. Traffic prediction accuracy.

Number of data	Traffic prediction accuracy (%)		
	ECMCRR-MPDNL (Proposed)	STCNet [1]	GNN-D [2]
500	90	86	82
1000	86	83	80
1500	88	84	81
2000	87	83	79
2500	85	80	77
3000	90	86	80
3500	93	89	84
4000	92	88	83
4500	93	86	84
5000	91	85	83

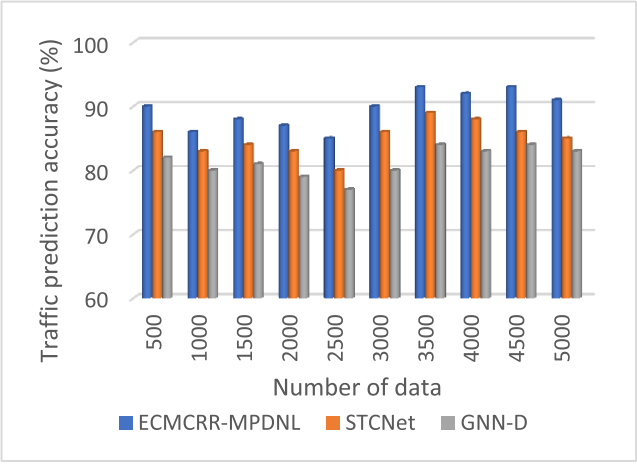

FIGURE 3. Graphical results of traffic prediction accuracy.

Totally ten accuracy results are obtained from the experimental evaluation with various inputs. The accuracy of the ECMCRR-MPDNL technique is compared with the existing results of existing methods. Then the average of ten results confirms that the traffic prediction accuracy is considerably improved by 5% compared to STCNet [1] and 10% compared to GNN-D [2] respectively.

B. FALSE-POSITIVE RATE

The False positive rate is also called a prediction error which referred to the ratio of a number of data incorrectly predicted as a possibility of traffic occurrences to the total number of data. The mathematical formula for calculating the false

TABLE 4. False positive rate.

Number of data	False positive rate (%)		
	ECMCRR-MPDNL (Proposed)	**STCNet [1]**	**GNN-D [2]**
500	10	14	18
1000	14	17	20
1500	12	16	19
2000	13	18	22
2500	15	20	23
3000	10	14	20
3500	7	11	16
4000	8	13	18
4500	7	14	16
5000	9	15	17

positive rate is given below,

$$PR_F = \left(\frac{Number\ of\ data\ incorrectly\ predicted}{n}\right) * 100 \tag{12}$$

where, PR_F denotes a false positive rate, n denotes the number of the data. The false-positive rate is measured in terms of percentage (%).

Sample calculation:

- **Proposed ECMCRR-MPDNL:** The number of data taken as input is 500 and the number of data incorrectly predicted is 50. Then the false positive rate is calculated as follows,

$$PR_F = \left(\frac{50}{500}\right) * 100 = 10\%$$

- **Existing STCNet:** The number of data taken as input is 500 and the number of data incorrectly predicted is 70. Then the false positive rate is calculated as follows,

$$PR_F = \left(\frac{70}{500}\right) * 100 = 14\%$$

- **Existing GNN-D:** The number of data taken as input is 500 and the number of data incorrectly predicted is 90. Then the false positive rate is calculated as follows,

$$PR_F = \left(\frac{90}{500}\right) * 100 = 18\%$$

Table 4 describes the experimental results of traffic prediction error i.e. false positive rate with respect to the number of data taken from the big dataset. The reported results evidently prove that the error rates 's' said to be minimized using the ECMCRR-MPDNL technique as compared to state-of-the-art methods. This is proved by the mathematical calculation. Let us consider 500 data for calculating the false positive rate. The 50 data are incorrectly predicted as traffic using the ECMCRR-MPDNL technique and their false positive rate is 10%. Similarly, 70 data and 90 data are incorrectly predicted, and their percentages are 14% and 18% respectively. Therefore, the various results of false-positive rates are obtained as shown in figure 4.

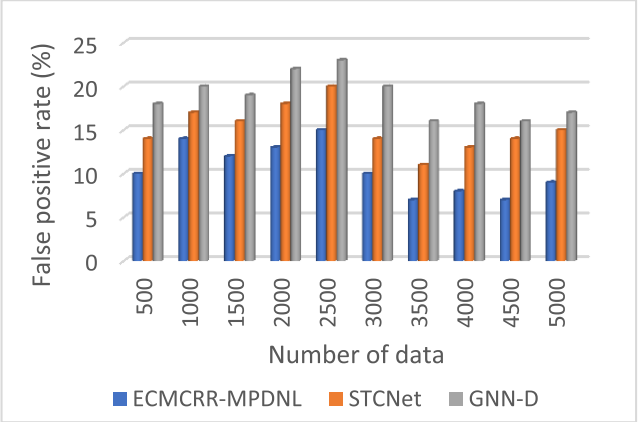

FIGURE 4. Graphical results false positive rate.

The graph illustrates the experimental results of the false-positive rate of traffic data prediction with respect to a number of data taken from the big dataset. The graphical results inferred that the network traffic prediction error is found to be minimized using the ECMCRR-MPDNL technique than the two existing techniques. This is due to the application of Ruzicka regression analysis. The analyzed results are given to the output layer for predicting the cellular network traffic. The activation function sets the threshold for the similarity value. If the similarity value is higher than the threshold is said to be a higher possibility of traffic occurrences at a particular space and time. Otherwise, the activation function provides the result as no traffic. Finally, the ECMCRR-MPDNL technique measures the man square error for observed results and actual results. The ECMCRR-MPDNL technique repeats the process until the error is minimized. This helps to minimize the incorrect traffic prediction in the cellular network. The average of the ten results evidently confirms that the ECMCRR-MPDNL technique minimizes the false positive rate by 32% compared to STCNet [1] and 45% compared to GNN-D [2] respectively.

C. TRAFFIC PREDICTION TIME

Traffic prediction time is defined as an amount of time taken by the algorithm to predict the network traffic with the given data. The prediction time is mathematically calculated as follows,

$$T_{TP} = n * time(predicting\ one\ data) \tag{13}$$

where, T_{TP} denotes a traffic prediction time, n represents a number of data. Traffic prediction time is measured in terms of milliseconds (ms).

Sample calculation:

- **Proposed ECMCRR-MPDNL:** The number of data taken as input is 500 and the time for predicting single data is $0.028ms$. Therefore, the overall prediction time

is computed as follows,

$$T_{TP} = 500 * 0.028ms = 14ms$$

- **Existing STCNet:** The number of data taken as input is 500 and the time for predicting single data is 0.034*ms*. Therefore, the overall prediction time is computed as follows,

$$T_{TP} = 500 * 0.034ms = 17ms$$

- **Existing GNN-D:** The number of data taken as input is 500 and the time for predicting single data is 0.038*ms*. Therefore, the overall prediction time is computed as follows,

$$T_{TP} = 500 * 0.038ms = 19ms$$

Table 5 shows the performance analysis of the prediction time of three different techniques and the number of data taken in the range from 500 to 5000. From the above-tabulated results, the traffic prediction time of the proposed ECMCRR-MPDNL technique is decreased. The tabulated results are plotted in figure 5.

TABLE 5. Traffic prediction time.

Number of data	Traffic prediction time (ms)		
	ECMCRR-MPDNL (Proposed)	**STCNet [1]**	**GNN-D [2]**
500	14	17	19
1000	20	24	27
1500	23	30	33
2000	28	34	36
2500	33	36	40
3000	36	41	44
3500	40	44	46
4000	44	46	50
4500	47	50	53
5000	50	53	58

Figure 5 depicts the comparison results of traffic prediction time versus a number of data. As shown in the figure, while increasing the number of data, the traffic prediction time also increased for all three techniques. But comparatively, traffic prediction time gets minimized using the ECMCRR-MPDNL technique. This significant improvement is achieved by applying a clustering concept in the deep multilayer precepted neural learning technique. The number of data is taken from the big dataset for partitioned into different groups to minimize the time complexity of the traffic prediction. The proposed expected conditional maximization clustering techniques groups the spatial data into different clusters based on the log-likelihood function. Then the clustering results are transferred into the next hidden layers for analyzing the data in the cluster with the traffic patterns. This process takes minimum time for predicting the traffic in a cellular network.

Let us consider 500 data in the first iteration for conducting the experiments. The ECMCRR-MPDNL technique

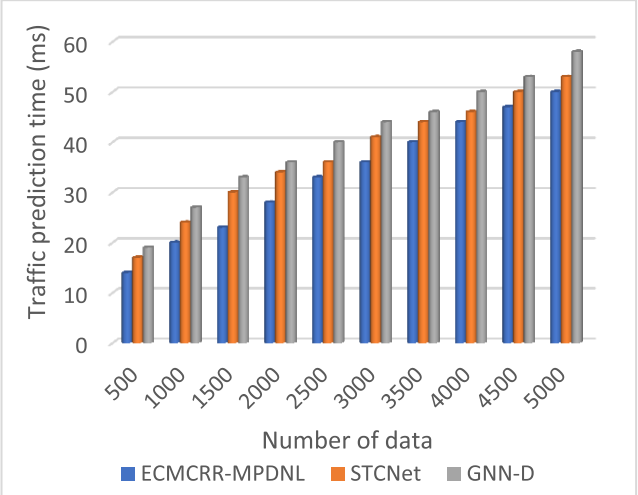

FIGURE 5. Graphical results of traffic prediction time.

utilizes 14ms of time for predicting the network traffic. The STCNet [1] and GNN-D [2] technique consumed 17*ms* and 19*ms* for predicting the cellular network traffic. Similarly, nine remaining iterations are carried out with various ranges of input data. The average of ten various experimental results of the proposed technique is compared to the existing methods. The compared results prove that the traffic prediction time is minimized by 12% and 19% when compared to the two existing techniques.

The above-discussed results confirm that the ECMCRR-MPDNL technique performs cellular network traffic prediction with higher accuracy and minimum time as well as error rate than the two state-of-the-art methods.

VI. CONCLUSION AND FUTURE SCOPE

An efficient technique called ECMCRR-MPDNL is introduced to achieve higher network traffic prediction accuracy and minimal time with the big cellular data. The Multilayer precepted deep neural learning concept collects the data from the big datasets and given to the input layer. Then the input data are learned with two hidden layers. In the first hidden layer, the clustering concept minimizes the time complexity of the traffic prediction by dividing the total network into different groups using maximum likelihood probability distribution. Then the regression function analyzes the data within the cluster using Ruzicka similarity measure at the second hidden layer. The learned data are transferred into the output layer to predict the network traffic using activation function with minimum error. This helps to improve the prediction accuracy and minimizes the false positive rate. An experiment is conducted using big datasets with different parameters such as traffic prediction accuracy, false positive rate and prediction time. The discussed result clearly shows that ECMCRR-MPDNL improves the traffic prediction accuracy and minimizes the prediction time as well as the false-positive rate when compared to the state-of-the-art methods. Future scope, there still exist some issues to be addressed. The major

challenge for the application-level traffic modeling prediction services continually blossom and emerge, Furthermore, it is still interesting to investigate how to leverage the additional information (e.g., inter-service relevancy) to further optimize the proposed framework.

REFERENCES

[1] C. Zhang, H. Zhang, J. Qiao, D. Yuan, and M. Zhang, "Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1389–1401, Jun. 2019.

[2] X. Wang, Z. Zhou, F. Xiao, K. Xing, Z. Yang, Y. Liu, and C. Peng, "Spatio-temporal analysis and prediction of cellular traffic in metropolis," *IEEE Trans. Mobile Comput.*, vol. 18, no. 9, pp. 2190–2202, Sep. 2019.

[3] Z. Liu, R. Wang, and D. Tang, "Extending labeled mobile network traffic data by three levels traffic identification fusion," *Future Gener. Comput. Syst.*, vol. 88, pp. 1–13, Nov. 2018.

[4] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin, "Understanding mobile traffic patterns of large scale cellular towers in urban environment," *IEEE/ACM Trans. Netw.*, vol. 25, no. 2, pp. 1147–1161, Apr. 2017.

[5] S. A. Abdulkarim and I. A. Lawal, "A cooperative neural network approach for enhancing data traffic prediction," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 25, pp. 4746–4756, Dec. 2017.

[6] S. Zhao, S. Chen, Y. Sun, Z. Cai, and J. Su, "Identifying known and unknown mobile application traffic using a multilevel classifier," *Secur. Commun. Netw.*, vol. 2019, pp. 1–11, Jan. 2019.

[7] R. Li, Z. Zhao, J. Zheng, C. Mei, Y. Cai, and H. Zhang, "The learning and prediction of application-level traffic data in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3899–3912, Jun. 2017.

[8] Q. T. Tran, L. Hao, and Q. K. Trinh, "Cellular network traffic prediction using exponential smoothing methods," *J. Inf. Commun. Technol.*, vol. 18, no. 1, pp. 1–18, Jan. 2019.

[9] C. Qiu, Y. Zhang, Z. Feng, P. Zhang, and S. Cui, "Spatio-temporal wireless traffic prediction with recurrent neural network," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 554–557, Aug. 2018.

[10] Y. Zhang, J. Li, Y. Li, D. Xu, M. Ahmed, and Y. Li, "Cellular traffic offloading via link prediction in opportunistic networks," *IEEE Access*, vol. 7, pp. 39244–39252, 2019.

[11] K. Zhang, G. Chuai, W. Gao, X. Liu, S. Maimaiti, and Z. Si, "A new method for traffic forecasting in urban wireless communication network," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, pp. 1–12, Mar. 2019.

[12] L. Nie, X. Wang, L. Wan, S. Yu, H. Song, and D. Jiang, "Network traffic prediction based on deep belief network and spatiotemporal compressive sensing in wireless mesh backbone networks," *Wireless Commun. Mobile Comput.*, vol. 2018, pp. 1–10, Jan. 2018.

[13] J. Feng, X. Chen, R. Gao, M. Zeng, and Y. Li, "DeepTP: An end-to-end neural network for mobile cellular traffic prediction," *IEEE Netw.*, vol. 32, no. 6, pp. 108–115, Nov. 2018.

[14] L. Chen, D. Yang, D. Zhang, C. Wang, J. Li, and T.-M.-T. Nguyen, "Deep mobile traffic forecast and complementary base station clustering for C-RAN optimization," *J. Netw. Comput. Appl.*, vol. 121, pp. 59–69, Nov. 2018.

[15] K. Sultan, H. Ali, A. Ahmad, and Z. Zhang, "Call details record analysis: A spatiotemporal exploration toward mobile traffic classification and optimization," *Information*, vol. 10, no. 6, pp. 1–17, 2019.

[16] R. Shinkuma, Y. Tanaka, Y. Yamada, E. Takahashi, and T. Onishi, "User instruction mechanism for temporal traffic smoothing in mobile networks," *Comput. Netw.*, vol. 137, pp. 17–26, Jun. 2018.

[17] E. M. R. Oliveira, A. C. Viana, K. P. Naveen, and C. Sarraute, "Mobile data traffic modeling: Revealing temporal facets," *Comput. Netw.*, vol. 112, pp. 176–193, Jan. 2017.

[18] C. Zhang, H. Zhang, D. Yuan, and M. Zhang, "Citywide cellular traffic prediction based on densely connected convolutional neural networks," *IEEE Commun. Lett.*, vol. 22, no. 8, pp. 1656–1659, Aug. 2018.

[19] A. Ozovehe, "Mobile soft switch traffic prediction using polynomial neural networks," *Eur. J. Eng. Res. Sci.*, vol. 3, no. 7, pp. 22–27, 2018.

[20] Y. Yur, M. F. Akay, and F. Abut, "Short term voice traffic forecast in 3G/UMTS networks using machine learning and statistical methods," *Int. J. Adv. Electron. Comput. Sci.*, vol. 3, no. 10, pp. 108–111, 2016.

[21] M. F. Iqbal, M. Zahid, D. Habib, and L. K. John, "Efficient prediction of network traffic for real-time applications," *J. Comput. Netw. Commun.*, vol. 2019, pp. 1–11, Feb. 2019.

[22] S. Goudarzi, M. Kama, M. Anisi, S. Soleymani, and F. Doctor, "Self-organizing traffic flow prediction with an optimized deep belief network for Internet of vehicles," *Sensors*, vol. 18, no. 10, p. 3459, Oct. 2018.

[23] N. G. Polson and V. O. Sokolov, "Deep learning for short-term traffic flow prediction," *Transp. Res. C, Emerg. Technol.*, vol. 79, pp. 1–17, Jun. 2017.

[24] S. Hakak, W. Z. Khan, G. A. Gilkar, M. Imran, and N. Guizani, "Securing smart cities through blockchain technology: Architecture, requirements, and challenges," *IEEE Netw.*, vol. 34, no. 1, pp. 8–14, Jan. 2020.

[25] M. A. Kaljahi, P. Shivakumara, S. Hakak, M. Y. I. Idris, M. H. Anisi, and D. Rajan, "Saliency-based bit plane detection for network applications," *Multimedia Tools Appl.*, vol. 1, pp. 1–19, Mar. 2020, doi: 10.1007/s11042-020-08741-9.

[26] F. Al-Turjman, "Intelligence and security in big 5G-oriented IoNT: An overview," *Future Gener. Comput. Syst.*, vol. 102, pp. 357–368, Jan. 2020.

[27] I. Mehmood, A. Ullah, K. Muhammad, D.-J. Deng, W. Meng, F. Al-Turjman, M. Sajjad, and V. H. C. de Albuquerque, "Efficient image recognition and retrieval on IoT-assisted energy-constrained platforms from big data repositories," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9246–9255, Dec. 2019.

[28] H. B. Salameh, S. Otoum, M. Aloqaily, R. Derbas, I. A. Ridhawi, and Y. Jararweh, "Intelligent jamming-aware routing in multi-hop IoT-based opportunistic cognitive radio networks," *Ad Hoc Netw.*, vol. 98, Mar. 2020, Art. no. 102035.

[29] I. Al Ridhawi, Y. Kotb, M. Aloqaily, Y. Jararweh, and T. Baker, "A profitable and energy-efficient cooperative fog solution for IoT services," *IEEE Trans. Ind. Informat.*, vol. 16, no. 5, pp. 3578–3586, May 2020.

[30] Y. Kotb, I. Al Ridhawi, M. Aloqaily, T. Baker, Y. Jararweh, and H. Tawfik, "Cloud-based multi-agent cooperation for IoT devices using workflow-nets," *J. Grid Comput.*, vol. 17, no. 4, pp. 625–650, Dec. 2019.

VENKATA SUBBARAJU DOMMARAJU (Member, IEEE) was born in Rajampet, India, in January 1992. He received the Bachelor of Technology (B.Tech.) degree in information technology from Jawaharlal Nehru Technological University Anantapur, Anantapur, India, in 2013, and the Master of Science (M.Sc.) degree in computer science from Northwestern Polytechnic University, in 2015. He is currently pursuing the Ph.D. degree in information technology with the University of the Cumberland, Williamsburg, KY, USA. He joined Corporate Pharmacy Company as a Programmer Analyst, USA, to gain practical knowledge and experience in information technology and provided several solutions to complex products and manufacturing infrastructure, from 2015 to 2018. His research interests include big data, data science in artificial intelligence and machine learning, the Internet of Things (IoT), databases, information retrieval, data mining, data analytics, medical research, manufacturing products, visual analytics, exploring methods in data science and big data technologies more sustainable, cost effective, and secure through extensive research and analysis on today's new technologies.

KARTHIK NATHANI (Member, IEEE) was born in Ongole, India, in August 1991. He received the Bachelor of Technology (B.Tech.) degree in electronics and communication engineering from Jawaharlal Nehru Technological University Hyderabad, India, in 2012, and the Master of Science degree in electrical engineering from Wright State University, Dayton, OH, USA, in 2014. He is currently pursuing the Ph.D. degree with the University of the Cumberland, in 2019. He worked as a Consultant to clients in the field of medical research. He was a Programmer Analyst with the State Department of Education and Commercial Insurance Providers, from 2015 to 2018, provides the best solutions to be unique and secure in current leading digital technologies. His research interests include internet-related services and products, information and communication technology develops the attributes required to successfully identify, investigate, and resolve problems and opportunities in today's IT industry.

USMAN TARIQ received the Ph.D. degree from Ajou University, South Korea. He led the design of a global data infrastructure simulator modeling, to evaluate the impact of competing architectures on the performance, availability, and reliability of the system for the industrial IoT infrastructure. He is currently an Associate Professor with the College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University. His international collaborations/collaborators, include but not limited to, such as NYIT, Ajou University, PSU, University of Sherbrooke, COMSATS, NUST, UET, the National Security Research Institute (NSR), Embry-Riddle Aeronautical University, Korea University, Manchester Metropolitan University, University of Bremen, and Virginia Commonwealth University. As a Network Security Theorist, his contributions towards addressing these challenges involve. His research interests include applied cyber security, advanced topics in the Internet of Things, health informatics, the theory of large complex networks, which includes network algorithms, stochastic networks, network information theory, and large-scale statistical inference.

FADI AL-TURJMAN received the Ph.D. degree in computer science from Queen's University, Kingston, ON, Canada, in 2011. He is currently a Full Professor and the Research Center Director with Near East University, Nicosia, Cyprus. He is also a Leading Authority in the areas of smart/intelligent, wireless, and mobile networks architectures, protocols, deployments, and performance evaluation. His publication history spans over 250 publications in journals, conferences, patents, books, and book chapters, numerous keynotes, and plenary talks at flagship venues. He has authored and edited more than 25 books about cognition, security, and wireless sensor networks' deployments in smart environments, published by Taylor and Francis, Elsevier, and Springer. He has received several recognitions and best papers awards at top international conferences. He also received the prestigious Best Research Paper Award from *Computer Communications Journal* (Elsevier), for the period of 2015 to 2018, and the Top Researcher Award from Antalya Bilim University, Turkey, in 2018. He has led a number of international symposia and workshops in flagship communication society conferences. He serves as an Associate Editor and the Lead Guest/Associate Editor for several well reputed journals, including the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS (IF 22.9) and *Sustainable Cities and Society* (Elsevier) (IF 4.7).

SURESH KALLAM (Member, IEEE) received the bachelor's and master's degrees from Jawaharlal Nehru Technological University Hyderabad and the Ph.D. degree from VIT University, Vellore, India. He was a Professor, the Division Head, and the Program Chair of the M.Tech. degree with the School of Computer Science Engineering, Galgotias University, Greater Noida, India. He was also a Foreign Faculty Member with the East China University of Technology, Jiangxi, China, and a Visiting Faculty Member with Jiangxi Normal University, China. He is currently an Associate Professor with the Department of Computer Science and Engineering, Sree Vidyanikethan Engineering College, Tirupati, and the Autonomous College, Rangampet. He has published over 35 national and international conference papers and 15 international journals. His research interests include the Internet of Things, big data, and high-performance computing. He received the Young Scientist Award, the Best Faculty Award, the Best Paper Award, in 2005, and the First Prize from the National Paper Presentation, in 2008.

PRAVEEN KUMAR REDDY M received the B.Tech. degree in CSE from JNT University and the M.Tech. degree in CSE from the Vellore Institute of Technology, Vellore, India. He has served with IBM and Alcatel-Lucent, as a Senior Software Engineer. He was a Visiting Professor with the Guangdong University of Technology, China, in 2019. He is currently an Assistant Professor with the School of Information Technology and Engineering, Vellore Institute of Technology. He has produced more than 15 international/national publications. His research interests include energy aware applications for the Internet of Things (IoT) and high-performance computing.

RIZWAN PATAN received the B.Tech. and M.Tech. degrees from Jawaharlal Nehru Technological University Anantapur, India, in 2012 and 2014, respectively, and the Ph.D. degree in computer science and engineering from the School of Computer Science and Engineering, VIT University, Vellore, India, in 2017. He was a former Assistant Professor with the School of Computing Science and Engineering, Galgotias University, India, from 2017 to 2019. He has been an Assistant Professor with the Department of Computer Science and Engineering, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, India, since 2019. He has published reputed 20 SCI journals and ten free Scopus indexed journals. He has also presented paper in National/International Conferences, published book chapters in CRC Press, IGI Global, Elsevier, and Edited as books. He holds over ten Indian patents and one USA patent. He received the Award from the World Research Council and the American Medical Council in the title of Innovative Researcher on Big Data and IoT, in 2019. He serves as a Guest Editor for the *International Journal of Grid and Utility Computing* (Inderscience), *Recent Patents on Computer Science*, *Informatics in Medical Unlocked* (Elsevier), and *Neural Computing and Applications* (Springer).