Deep Learning based Traffic Prediction for Resource Allocation in Multi-Tenant Virtualized 5G networks

Preety Rebari
Computer Science and Engineering
National Insitute of Technology Warnagal
rebaripreety@gmail.com

Balaprakasa Rao Killi
Computer Science and Engineering
National Insitute of Technology Warnagal
bsprao@nitw.ac.in

***Abstract*—Network traffic has changed significantly since the introduction of 5G technology. The massive flow of connected devices as well as new applications are causing a variety of traffic patterns, quality of service (QoS) requirements and scalability challenges. The solution to these problems is network slicing which enables operators to distribute resources and set network characteristics per slice by constructing multiple virtual network slices within the shared 5G infrastructure. Resources are reserved for each slice for some time period. Traffic prediction with high accuracy is of great importance in the dynamic environment of 5G networks for resource planning and scheduling as well as for reliable and effective transmission of network data. In this paper, We generate user requests in a slice having arrival and departure rates using Poisson distribution. We did a comparative analysis of traffic prediction and resource allocation using deep learning models such as Long Short Term Memory(LSTM), Bidirectional LSTM(BiLSTM), Stacked LSTM and Gated Recurrent Unit(GRU).**

***Index Terms*—5G Network, Network Slicing, Traffic Prediction, Resource Allocation, LSTM, Stacked LSTM, Bidirectional LSTM, GRU.**

I. Introduction

5G Network is the fifth generation wireless cellular technology succeeding the fourth generation (4G) LTE network technology. The major purpose of 5G is to provide faster and more stable wireless communication. It is capable of handling significantly higher data rates and with much lower latency than 4G. Data rates in 5G are expected to be substantially higher than 4G. 5G provides faster downloads and uploads of huge files and smoother streaming of high-quality video material. 5G networks are designed to have substantially lower latency than 4G networks which implies that data may be delivered between devices much faster. This is especially important for real-time applications such as remote surgery, driverless cars and virtual reality. Overall, 5G network technology is predicted to significantly increase mobile network connectivity which will enable new and innovative applications and use cases that earlier generations of mobile network technology were unable to support.

Three different categories of services that the 5G network offers are enhanced mobile broadband(eMBB), massive machine type communication(mMTC) and ultra-reliable low latency communication(uRLLC). eMBB focuses on offering end-users faster and more reliable mobile broadband. eMBB services are intended to transmit high-speed data to smartphones, tablets and other mobile devices at up to ten times the speed of 4G. Video streaming, internet gaming and virtual and augmented reality are all examples of EMBB applications [1]. uRLLC focuses on providing ultra-reliable, low-latency communication for mission-critical applications. uRLLC services are meant to provide quick and dependable communication for Remote surgery, driverless cars and industrial automation which are all examples of real-time applications. mMTC is focused on connecting a large number of devices, such as IoT gadgets and sensors. mMTC services are intended to accommodate a large density of connected devices while consuming little power and costing little money. Smart cities, smart homes and industrial automation are some of the applications for mMTC [2].

Network slicing in 5G is a technique that allows network operators to construct several virtual networks, also known as network slices, using a single physical 5G network infrastructure. Each network slice is tailored to specific application or user requirements such as low latency for driverless vehicles, high bandwidth for streaming video or high dependability for industrial automation. Network slicing divides the 5G network into many logical portions that may be designed and controlled individually. Each network slice has its own set of resources, including computing power, storage, and bandwidth which can be allocated and optimized dynamically based on the demands of the applications or users that use that slice. Network slicing enables network operators to provide customized services to various consumers and applications while also maximizing network resource use. It also allows for the quick deployment of new services and applications since network slices may be supplied and updated quickly without affecting other slices or overall network performance. Overall, 5G network slicing [3] is a strong tool for network operators to unleash the full potential of 5G technology and allows for a broad spectrum of applications from consumer-oriented to mission-critical industrial.

To address the enormous traffic load which resulted from massive diverse data that increase the deployment and distribution of 5G network resources in major cities, and improve the intelligence and dependability of traffic management. It is of utmost importance to estimate traffic with high precision. The 5G network traffic is primarily time-series data, thus the prediction problem may be turned into a time-series prediction

forecasting task. As a result, many artificial intelligence (AI)-based machine learning (ML) and deep learning (DL) models have been devised for predicting mobile data traffic.

In this paper, the Poisson distribution is used to generate user requests in slices having arrival and departure rates. We distributed resources to each slice across a predetermined time frame known as a prediction window (PW) which consists of several short time steps. We predict the network data for the next PW from the values of previous and current PW using deep learning models mainly consisting of LSTM, Stacked LSTM, Bidirectional LSTM, and GRU.

The remaining sections are arranged as follows: Section 2 presents a review of the present state of the art in traffic prediction and resource allocation in 5G networks. In Section 3, we discuss problem definition and various deep learning models for traffic prediction. Section 4 showcases the experimental results and section 5 provides a comprehensive summary of our work.

II. Related Work

The incorporation of software-defined networking (SDN) and network function virtualization (NFV) technologies with network slicing in 5G is presented in [4]. The combination of SDN and NFV allows for dynamic and efficient management of network slices. The paper [5] investigates the possibilities of applying network slicing in various vertical industries, including transportation, healthcare, and smart cities. Authors in [6] proposed an intent-based networking (IBN) approach that can separate the core network and radio access network domains to address the challenges in deploying and managing 5G networks. In [7], a network slice broker tries to efficiently allocate network resources to fulfill the unique needs of various users and applications. It manages network resources and slices the network into many virtual networks, each tailored to a unique user/application requirement using an SDN architecture.

The paper [8] presents a study on cellular data traffic prediction using recurrent neural networks (RNNs) with a focus on LSTM networks. The research demonstrates that LSTM models outperform both the Statistical autoregressive integrated moving average (ARIMA) model and conventional feed-forward neural networks (FFNN) in terms of prediction accuracy. The work in [9] uses LSTM neural networks to detect patterns in traffic data and generate accurate predictions. The results suggest that the proposed model outperforms traditional prediction models like ARIMA and linear regression. In [10], the authors aim to predict traffic flow in real-time which can help operators to optimize network resource allocation and manage network congestion. They collected the actual traffic data from a live network and compared effectiveness of gated recurrent unit (GRU) and LSTM against artificial neural networks (ANNs) and ARIMA. In paper [11] hybrid approach of statistics and deep learning was proposed to predict traffic in 5G networks. This approach provides true predictions and is helpful to network operators in planning a network resource allocation. The paper [12] 5G networks traffic is predicted using SDN and fusion learning. The authors collect network traffic data from multiple sources such as switches and routers and generate a unified view of the network by making use of SDN. This entire network view aids in forecasting future network traffic patterns. A machine learning-based methodology for mobile data traffic forecasting in 5G cellular networks was proposed in paper [13]. The LSTM model outperforms random forest(RF) and support vector regression(SVR) indicating its effectiveness for improving resource allocation and network optimization in 5G networks. In [14], a smoothed LSTM is introduced as a novel technique for 5G traffic prediction with enhanced accuracy and stability over typical LSTM models.

The paper [15] provides an overview of the fundamental concepts, practical uses, and challenges associated with the implementation of deep reinforcement learning(DRL) techniques in 5G networks and moreover. The work in [16] highlights the capabilities of DRL methods in resource allocation for 5G communications. The suggested approach exhibits promising results in adapting to changing network conditions and improving system performance. The authors in [17] proposed a technique that integrates DRL and network slicing to optimize the distribution of resources in networks that experience fluctuations in traffic patterns and have diverse resource needs. By employing this method, network slicing can be effectively implemented in dynamic scenarios, leading to efficient resource utilization and improved network performance. In paper [18] authors employed a deep Q-network (DQN) architecture, which integrates deep learning and reinforcement learning, to optimize the allocation of resources in network slices. It aims to enhance resource utilization, elevate service quality, and reduce the operational expenses associated with network slicing. The authors in [19] presented a deep reinforcement learning technique that incorporates discrete normalized advantage functions. This suggested method aims to maximize the allocation of resources to improve overall network performance by leveraging deep learning techniques. Incorporating IBN with 5G networks provides flexible, quick, and programmable network slicing, enabling service providers to effectively distribute network resources and satisfy the various demands of various applications and services [20].

III. Methodology

A. Problem Definition

We generate user requests in slice m using the Poisson distribution having departure and arrival rates of $\mu[m]$ and $\lambda[m]$, respectively. Resources are allocated to each slice during a specified time period known as the prediction window(PW) which consists of many small time intervals. More particularly, the resource allocation occurs at the start of the next PW. Note that the duration of the PW and the time intervals within it can be dynamically altered to accommodate the needs of mobile network slices such as the characteristics of 5G new radio (NR).

By analyzing the $r_m(t)$ values of the previous and current PWs, we calculate the $\bar{r}_m(t)$ for the next PW to reduce the MSE value for the predicted values at each prediction time.

In order to address this issue, we utilized Deep Learning Models such as LSTM, Stacked LSTM, Bidirectional LSTM and GRU. These are recurrent neural networks that can learn the dependencies within a sequential data sequence so as to anticipate a time series problem.

In this technique, the input data for deep learning models abbreviated as D_{input}, comprises the actually required resources ($A_m(t)$) from the previous and current PW_s for each slice at every time step. It is denoted as $\{A_m$(t-2PW),...,A_m(t-PW),...,A_m(t)$\}$. This input data is used in the next PW to forecast the allocated resources for the m^{th} slice at time step t. The slice isolation degree is determined by the confidence level χ.

The isolation of individual slices implies maintaining a consistent allocation of resources without changes over a defined timeframe. This is essential for ensuring service quality by preventing traffic load variations in one slice from impacting others. To achieve this, it is necessary to keep resources for future users of each slice, avoiding frequent resource reconfiguration. As a result, each slice is assigned a dedicated portion of the total resources (θ) exclusively for a certain period, while also having the flexibility to utilize Residual resources if additional capacity is required.

Let the predicted resource allocation for m^{th} slice in the next PW be represented by $P_{predicted} = \{p_m(t+1), ..., p_m(t+PW)\}$.The sample mean and standard deviation of $P_{predicted}$ are computed as $\bar{p}_m = \frac{1}{PW}\sum_{k=1}^{PW} p_m(t+k)$ and $\sigma(p_m) = \sqrt{\frac{1}{PW}\sum_{k=1}^{PW}(p_m(t+k) - \bar{p}_m)^2}$ respectively. During simulation, the confidence level χ, which can be adjusted dynamically for each slice, ranges between 0 and 1. Thus, the confidence interval can be determined as $r_m(t) \epsilon \bar{p}_m \pm Z_{(\frac{1-\chi}{2})}.\frac{\sigma(p_m)}{\sqrt{PW}}, t\epsilon[t, t+D_\Delta]$, whereas D_Δ denotes the PW. This interval specifies the amount of resources to be allocated ($r_m(t)$) to slice m in the next PW. To guarantee service quality, slice isolation, and compliance with traffic service level agreements (SLAs), the maximum value of the prediction interval is employed to allocate resources for slice m in the upcoming PW. This allocation takes into consideration the association between predicted values and errors.

B. Predictive Deep Learning Models

1) Long short-term memory(LSTM): A neural network architecture called an LSTM was created to manage long-term dependencies in sequential data. It has proven to be highly effective in tasks such as speech recognition, natural language processing and time-series prediction. To accomplish their efficacy LSTMs employ a memory cell and a collection of gates that govern the flow of information into and out of the cell as shown in Fig. 1. The gates choose which information to keep and which to discard, while the memory cell records and updates the current state of the network. The gates used in an LSTM are the input gate, forget gate and output gate. The input gate controls how much new information enters the cell while the forget gate controls how much old information is removed. The output gate controls how much data is transferred to the final output or the next layer of the network.

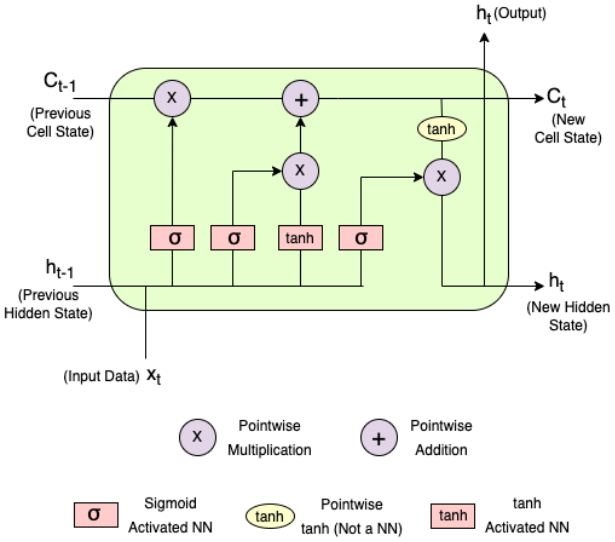


Fig. 1: LSTM

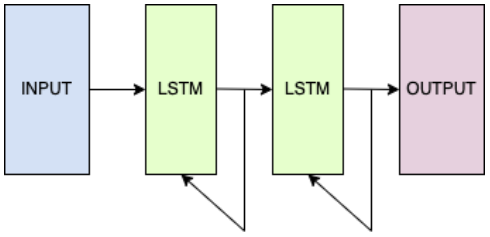


Fig. 2: Stacked LSTM

2) Stacked LSTM: As seen in Fig. 2, a stacked LSTM is a particular form of neural network design that comprises of many layers of LSTM units. Each layer processes the input data at different levels of abstraction and the output of one layer is utilised as the input for the next layer. In other words, after receiving the input data, the first layer produces a set of output values, which are then passed on to the second layer. A fully connected layer receives the output of the final layer and uses it to create the network's ultimate output. For problems where the model must learn hierarchical representations of the input data, stacked LSTMs are especially helpful. The model can learn more complex associations and recognise longer-term dependencies in the input data by piling up numerous LSTM layers on top of one another.

3) Bidirectional LSTM: The forward and backward information flow of a sequence through an LSTM network are combined in a neural network architecture called a BiLSTM. A BiLSTM processes the input sequence in both forward and backward directions using two independent LSTM layers.The outputs of both layers are then concatenated at each time step to form the final output which incorporates data from both directions and is represented in Fig. 3 A BiLSTM can capture both forward and backward context by processing the input sequence in both ways which can result in predictions that are more accurate.

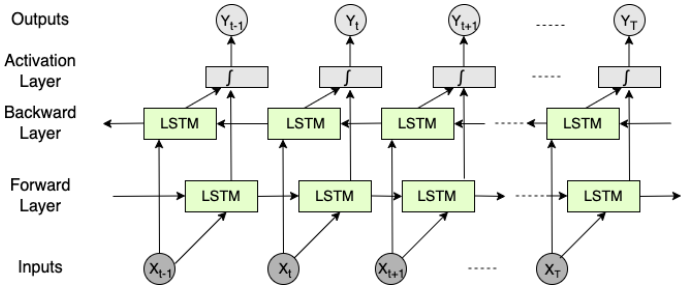


Fig. 3: Bidirectional LSTM

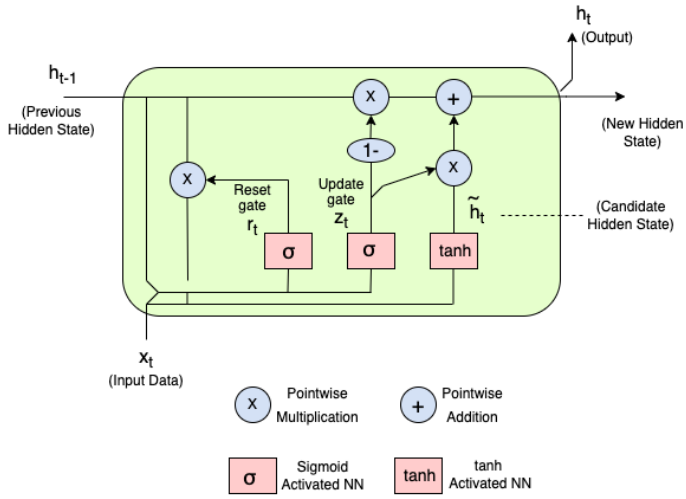


Fig. 4: GRU

4) Gated recurrent unit(GRU): It is a form of RNN architecture. It is similar to LSTM in that it is meant to capture long-term relationships in sequential data, but it has a simpler architecture and fewer parameters. GRUs are made up of recurrent neural network units that are equipped with gates that regulate the flow of information through the network. The gates in a GRU include an update gate and a reset gate which control which information is passed from the previous time step to the current time step as shown in Fig. 4. A GRU's update gate regulates how much of the previous hidden state is transferred to the current time step, while the reset gate decides how much of the current input is mixed with the prior hidden state. These gates enable the GRU to preserve or reject prior time step information according to its relevance to the current time step.

IV. Experimental Results and Discussions

In this section, we compare the performance of four deep learning models for resource requirement prediction: LSTM, stacked LSTM, Bidirectional LSTM, and GRU.

A. Simulation Environment

In this simulation, there are 10 slices, 200 RAN RBs, and a Prediction Window (PW) of 600 seconds. The λ values assigned to the 10 slices are [0.7, 0.75, 0.6, 0.75, 0.65, 0.75, 0.6, 0.8, 0.7, 0.75], while the corresponding μ values are [0.75, 0.85, 0.75, 0.85, 0.75, 0.85, 0.76, 0.85, 0.75, 0.85]. It should be noted that a PW consists of numerous small time steps. Specifically, large-scale resource allocation occurs at the start of the next PW, whereas small-scale resource allocation occurs at each time step. In our simulations, the total number of time steps is 60,000 and the PW size is 600. The simulations were carried out in Python utilizing the Keras and Tensorflow packages.

B. Dataset and Preprocessing

For each second in a slice, we have used a Poisson distribution to generate the users' arrival and departure requests having departure and arrival rates μ and λ respectively. From the arrival and departure user requests, we then calculated the ongoing requests for every second. Later used this data as input data to the deep learning models.

C. Establish deep learning models

The input and output sequences are split into training and testing sets, in which 80% of the data is used as the training set and 20% as the testing set. All the Deep learning model layers consist of 100 units followed by a dense layer. The models are compiled with the mean squared error as the loss function and the Adam optimizer. During training models, 20% of training set is used as validation set. Models are trained for 100 epochs with a batch size of 1.

D. Performance comparison of traffic prediction models

In Fig. 5, we first evaluate the prediction of resource requirement over 12000 timesteps for LSTM (Fig. 5a), Stacked LSTM (Fig. 5b), Bidirectional LSTM (Fig. 5c) and GRU (Fig. 5d)respectively and compared the results with the actual resource requirement. The blue line indicates the actual traffic data while the red line indicates the predicted traffic data. The experiment demonstrates that the predicted values for the GRU Model accurately represent the actual traffic values.

E. Evaluation Index

Mean squared error (MSE) is a popular metric for assessing the performance of deep learning models. MSE is calculated by comparing the model's predicted values to their corresponding true values. The squared differences across all samples in the dataset are then averaged to determine the MSE.

Equation 1 shows the mathematical expressions for the evaluation formula, where y signifies the predicted values and f signifies the actual values. A reduced MSE value indicates that the model's predictions are more accurate, which implies a higher level of performance.Conversely, a higher MSE value signifies larger prediction errors and lower performance.

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - f_t)^2 \quad (1)$$

Fig. 6 shows the mean squared error in terms of the number of epochs for LSTM (Fig. 6a), Stacked LSTM (Fig. 6b),

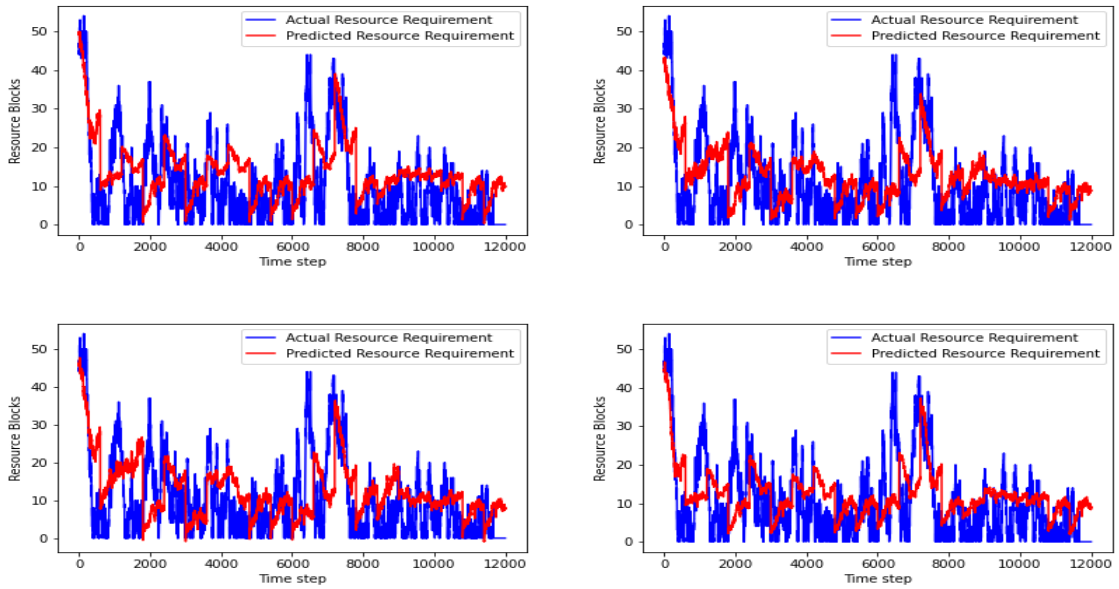
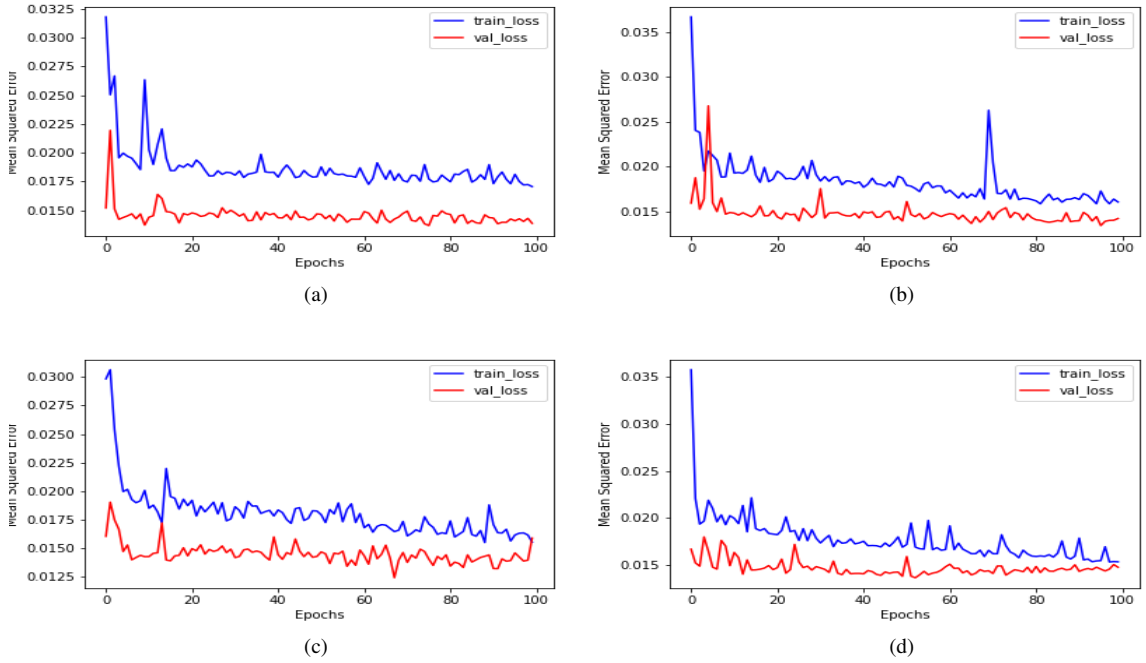


Fig. 5: Resource requirement prediction for Poisson Arrivals and Departures. (a) LSTM. b) Stacked LSTM c) BiLSTM d) GRU



(a) (b)

(c) (d)

Fig. 6: Training and Validation Loss for Poisson Arrivals and Departures. (a) LSTM. b) Stacked LSTM c) BiLSTM d) GRU

Bidirectional LSTM (Fig. 6c), and GRU (Fig. 6d) respectively for Poisson distribution data.

Fig. 7 shows the Mean Squared Error Comparison of LSTM, Stacked LSTM, BiLSTM, and GRU methods for Poisson distribution data. The experiment demonstrates that the MSE loss for the GRU Model is 0.0682 which is less as compared to the other models.

V. CONCLUSION

In this paper, We generated user requests in a slice having arrival and departure rates using Poisson distribution. We predict the resource allocation for the next PW from the values of previous and current PW. We did a comparative analysis of traffic prediction and resource allocation on traffic generated through Poisson distribution using deep learning models such as LSTM, Stacked LSTM, Bidirectional LSTM, and GRU. From the comparative analysis, we conclude that the GRU

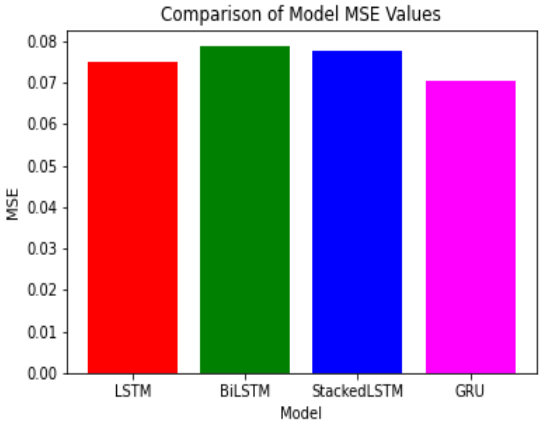


Fig. 7: Mean Squared Error Comparison of LSTM, Stacked LSTM, BiLSTM, GRU methods for Poisson Arrivals and Departures.

model performs better than the other three models on traffic generated through Poisson distribution.

REFERENCES

[1] Sengupta, Ayan, Alberto Rico Alvarino, Amer Catovic, and Lorenzo Casaccia. "Cellular terrestrial broadcast—Physical layer evolution from 3GPP release 9 to release 16." IEEE Transactions on Broadcasting 66, no. 2 (2020): 459-470.

[2] Pokhrel, Shiva Raj, Jie Ding, Jihong Park, Ok-Sun Park, and Jinho Choi. "Towards enabling critical mMTC: A review of URLLC within mMTC." IEEE Access 8 (2020): 131796-131813.

[3] Chen, Qiang, Xiaolei Wang, and Yingying Lv. "An overview of 5G network slicing architecture." In AIP conference proceedings, vol. 1967, no. 1, p. 020004. AIP Publishing LLC, 2018.

[4] Lucena, Jose, Pablo Ameigeiras, Diego Lopez, Juan J. Ramos-Munoz, Javier Lorca, and Jesus Folgueira. "Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges." IEEE Communications Magazine 55, no. 5 (2017): 80-87.

[5] Li, Xin, Mohammed Samaka, H. Anthony Chan, Deval Bhamare, Lav Gupta, Chengcheng Guo, and Raj Jain. "Network slicing for 5G: Challenges and opportunities." IEEE Internet Computing 21, no. 5 (2017): 20-27.

[6] K. Abbas, M. Afaq, T. Ahmed Khan, A. Rafiq, and W.-C. Song, "Slicing the core network and radio access network domains through intent-based networking for 5G networks," Electronics, vol. 9, no. 10, 2020. [Online]. Available: https://www.mdpi.com/2079-9292/9/10/1710

[7] V. Sciancalepore, L. Zanzi, X. Costa-Perez, and A. Capone, "Onets: Online network slice broker from theory to practice," 2018, arXiv:1801.03484.

[8] Jaffry, Shan, and Syed Faraz Hasan. "Cellular traffic prediction using recurrent neural networks." In 2020 IEEE 5th International Symposium on Telecommunication Technologies (ISTT), pp. 94-98. IEEE, 2020.

[9] Selvamanju, E., and V. Baby Shalini. "Deep Learning based Mobile Traffic Flow Prediction Model in 5G Cellular Networks." In 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), pp. 1349-1353. IEEE, 2022.

[10] Tuna, Evren, and Alkan Soysal. "LSTM and GRU based Traffic Prediction Using Live Network Data." In 2021 29th Signal Processing and Communications Applications Conference (SIU), pp. 1-4. IEEE, 2021.

[11] Mohseni, Maryam, Soodeh Nikan, and Abdallah Shami. "AI-based Traffic Forecasting in 5G network." In 2022 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 188-192. IEEE, 2022.

[12] Selvi, K. Tamil, and R. Thamilselvan. "An intelligent traffic prediction framework for 5G network using SDN and fusion learning." Peer-to-Peer Networking and Applications 15, no. 1 (2022): 751-767.

[13] Selvamanju, E., and V. Baby Shalini. "Machine learning based mobile data traffic prediction in 5g cellular networks." In 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1318-1324. IEEE, 2021.

[14] Gao, Zihang. "5G traffic prediction based on deep learning." Computational Intelligence and Neuroscience 2022 (2022).

[15] Xiong, Zehui, Yang Zhang, Dusit Niyato, Ruilong Deng, Ping Wang, and Li-Chun Wang. "Deep reinforcement learning for mobile 5G and beyond: Fundamentals, applications, and challenges." IEEE Vehicular Technology Magazine 14, no. 2 (2019): 44-52.

[16] Tham, Mau-Luen, Amjad Iqbal, and Yoong Choon Chang. "Deep reinforcement learning for resource allocation in 5G communications." In 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1852-1855. IEEE, 2019.

[17] Koo, Jaehoon, Veena B. Mendiratta, Muntasir Raihan Rahman, and Anwar Walid. "Deep reinforcement learning for network slicing with heterogeneous resource requirements and time varying traffic dynamics." In 2019 15th International Conference on Network and Service Management (CNSM), pp. 1-5. IEEE, 2019.

[18] Li, Rongpeng, Zhifeng Zhao, Qi Sun, I. Chih-Lin, Chenyang Yang, Xianfu Chen, Minjian Zhao, and Honggang Zhang. "Deep reinforcement learning for resource management in network slicing." IEEE Access 6 (2018): 74429-74441.

[19] Qi, Chen, Yuxiu Hua, Rongpeng Li, Zhifeng Zhao, and Honggang Zhang. "Deep reinforcement learning with discrete normalized advantage functions for resource management in network slicing." IEEE Communications Letters 23, no. 8 (2019): 1337-1341.

[20] Abbas, Khizar, Muhammad Afaq, Talha Ahmed Khan, Adeel Rafiq, and Wang-Cheol Song. "Slicing the core network and radio access network domains through intent-based networking for 5g networks." Electronics 9, no. 10 (2020): 1710.