This article has been accepted for publication in IEEE Wireless Communications Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LWC.2022.3195517

Two-stage Traffic Load Prediction Based Resource Reservation for Sliced HSR Wireless Networks

Li Yan, Xuming Fang, *Senior Member, IEEE,* Yuguang Fang, *Fellow, IEEE,* Yi Li, Qing Xue,

***Abstract*—In this letter, we propose a two-stage traffic load prediction scheme for network slices (NSs) in high-speed railway (HSR) wireless networks, where in the first stage, the K-means algorithm is leveraged to cluster traffic flows, and in the second stage, the long-short term memory (LSTM) algorithm is applied to predict the traffic load. Based on the obtained traffic features (including traffic volume and user velocity) and the network radio resource characteristics (including coverage performance and capacity), we design a service-tailored resource reservation mechanism. Simulation results show that our proposed scheme can significantly improve the traffic load prediction accuracy to ensure the NS resource reservation performance.**

***Index Terms*—HSR wireless networks; network slicing; traffic load prediction; resource reservation; machine learning**

I. Introduction

To satisfy the diverse requirements of emerging abundant high-speed railway (HSR) mobile services during the evolution of HSR industry towards intelligentization, network slicing, which is used to enable isolated and service-tailored radio resource management, is envisioned as a promising technology in the fifth generation mobile communication system for railway (5G-R) [1]. Based on the primary transmission requirements, all the HSR applications are generally classified into three categories served by three network slices (NSs) [2], i.e., the enhanced massive broadband (eMBB) communication served by the mobile-broadband slice, the ultra reliable low-latency communication (URLLC) served by the mission-critical slice, and the massive machine type communication (mMTC) served by the machine-type slice. Nevertheless, even pertaining to the same service category, different HSR applications may have different traffic patterns. For instance, the text and the voice URLLC data have different traffic patterns. The event-triggered sensors and the periodically-updated sensors have different traffic patterns [3], although they both pertain to the mMTC. On-board passenger video services and track-side video surveillance, both of which belongs to the eMBB, have different mobility characteristics. In sliced wireless networks, to guarantee the resource availability, radio resources are reserved for each slice based on the traffic load prediction. In HSR, to ensure the transmission reliability, sufficient radio resources should be preferentially reserved for the mission-critical slice. However, excessive reservations lead to resource wastes. Consequently, to guarantee the radio resource availability and utilization, the traffic patterns within an NS should be carefully distinguished in order to improve the traffic load prediction accuracy.

In most existing works on resource reservation, the main objective is either to reduce the blocking rate for high-priority users, like handover users [4], or to lower the signaling overhead of resource allocation for services with strong periodicity, like voice over IP (VoIP) [5]. In 5G, as a new paradigm, network slicing achieves service-tailored resource managements, in which dedicated resources are reserved on demand for target services with the time granularity of NS life-cycles, so as to satisfy diverse transmission requirements. In [6], a two-timescale resource management scheme was proposed for 5G RAN network slicing, including long-term resource reservation and short-term intra-slice resource allocation. In [7], a hierarchical combinatorial auction mechanism for network slicing resource managements was proposed, where each tenant submits its bid to a centralized controller for a certain amount of resources. However, these works jointly formulated and solved the resource reservation and allocation, and hardly ensured the performance of the resource reservation which is largely determined by the traffic load prediction. In [8], deep learning, i.e., the long-short term memory (LSTM) algorithm, was employed to improve the cellular traffic load prediction accuracy. Nevertheless, the service categories were not considered. In this letter, we propose a two-stage traffic load prediction scheme to improve the traffic load prediction accuracy for high-speed railway wireless networks, where in the first stage, the K-means algorithm is applied to perform traffic flow clustering, while in the second stage, LSTM is applied to predict traffic volumes. Compared with existing works, our proposed scheme further distinguishes the traffic flows pertaining to the same NS. Moreover, the velocity of a user is considered in the traffic flow clustering phase, so as to facilitate the subsequent resource management under HSR. Simulation results demonstrate the effectiveness of our proposed scheme. Then, according to the user velocity and traffic volume obtained from the traffic load prediction, we design a service-tailored resource reservation mechanism to get a good match between user requests and radio resources for sliced HSR wireless networks.

Li Yan and Xuming Fang are with Key Lab of Information Coding & Transmission, Southwest Jiaotong University, Chengdu 610031, China (E-mails: liyan@swjtu.edu.cn, xmfang@swjtu.edu.cn). Yuguang Fang is with Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA (Email: fang@ece.ufl.edu). Yi Li is with China Academy of Railway Sciences, Beijing, China (Email: liyi_swjtu@foxmail.com). Qing Xue is with Chongqing University of Posts and Telecommunications, Chongqing, China (Email: xueq@cqupt.edu.cn).

The work was supported in part by NSFC under Grant 62101460, 62001071,and 62071393, NSFC High-Speed Rail Joint Foundation under Grant U1834210, Sichuan Provincial Applied Basic Research Project under Grant 2020YJ0218, Sichuan Provincial NSFC under Grant 2022NSFSC0893, and Project of China Railway Corporation under Grant P2020G004.

II. Two-stage Traffic Load Prediction

A. K-means based intra-NS traffic flow clustering

As aforementioned, all existing HSR applications can be roughly classified into three categories, and within each service category, the traffic patterns of different data flows may still be different. Consequently, to improve the traffic load prediction so as to realize accurate resource reservation, we propose a two-stage traffic load prediction scheme. In the first stage, based on the traffic features, we leverage the K-means algorithm to cluster the intra-NS traffic flows. Then, in the second stage, the traffic volume of each cluster is input to the LSTM algorithm for the final traffic load predictions. The design of our proposed two-stage traffic load prediction scheme is shown in Fig. 1, where for clarity the subscripts of machine-type slice, mission-critical and mobile-broadband slice are numbered as 1, 2 and 3, respectively, and M represents the number of observation windows used for prediction. Since the K-means and LSTM algorithms are typically used in clustering and prediction, respectively, we select them in our design. Nevertheless, our design can also be generalized to other algorithms.

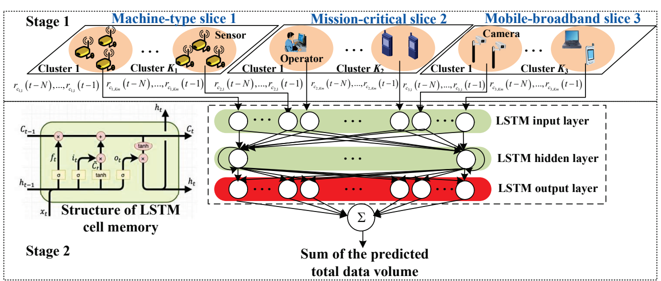


Fig. 1. The LSTM based traffic load prediction.

Without loss of generality, we assume M NSs with each NS serving U_m data flows, which are eventually classified into K_m clusters based on their traffic patterns. In this study, we define the number of data packets, the data volume per packet, and the mobility velocity of the data flow as the traffic features. The observation window is denoted as T_0. Note that here the mobility feature of data flows is the velocity of the user devices generating them. Since in 5G heterogenous networks, diverse resources have different mobility support, the mobility feature is considered in the clustering stage so as to instruct the subsequent resource reservation. Then, for data flow i in slice m, the traffic feature vector in the t-th observation window can be expressed as

$$\mathbf{s}_{m,i}\left(t\right) = \left[n_{m,i}(t), d_{m,i}(t), v_{m,i}(t)\right]. \tag{1}$$

where $n(t)$, $d(t)$ and $v(t)$ represent the number of packets, the data volume per packet, and the mobility velocity of data flows, respectively.

For slice m, the observed traffic samples of all U_m data flows within the t-th observation window form a sample set as

$$\mathbf{D}_m\left(t\right) = \left\{\mathbf{s}_{m,1}\left(t\right), \mathbf{s}_{m,2}\left(t\right), ..., \mathbf{s}_{m,U_m}\left(t\right)\right\}. \tag{2}$$

Suppose that all U_m data flows of slice m are eventually classified into K_m clusters, i.e.,

$$\mathbf{C}_m\left(t\right) = \left\{\mathbf{c}_{m,1}\left(t\right), \mathbf{c}_{m,2}\left(t\right), ..., \mathbf{c}_{m,K_m}\left(t\right)\right\}, \tag{3}$$

then, the optimization objective of the K-means clustering algorithm to minimize the total distance between the cluster members and the mean vectors of corresponding clusters can be formulated as [9]

$$\min E_m\left(t\right) = \sum_{i=1}^{K_m} \sum_{\mathbf{s}_{m,j}(t)\in \mathbf{c}_{m,i}(t)} \left\|\mathbf{s}_{m,j}\left(t\right) - \mathbf{m}_{m,i}\left(t\right)\right\|^2, \tag{4}$$

where $\mathbf{m}_{m,i}\left(t\right)$ represents the mean vector of cluster $\mathbf{c}_{m,i}\left(t\right)$ given below

$$\mathbf{m}_{m,i}\left(t\right) = \frac{1}{\left|\mathbf{c}_{m,i}\left(t\right)\right|} \sum_{\mathbf{s}_{m,j}(t)\in \mathbf{c}_{m,i}(t)} \mathbf{s}_{m,j}\left(t\right). \tag{5}$$

From the clustering results, we obtain the mean vectors of K_m traffic patterns, which can be employed as reference standards by network controllers to reserve NS resources. For instance, the traffic flows belonging to clusters with higher velocity should be assigned with radio resources with larger network coverage, which will be further discussed in detail in Section III. In the following subsection, by taking the traffic volumes after clustering as inputs, we design the LSTM algorithm for traffic load prediction.

B. LSTM based traffic load prediction

In the proposed two-stage traffic load prediction scheme as shown in Fig. 1, the K-means traffic flow clustering algorithm of the first stage aims to distinguish traffic patterns within NSs, while in the second stage the traffic loads of different clusters are taken as the input to the LSTM algorithm to predict the traffic loads. Based on the definition of traffic feature vector in (1), the traffic loads of cluster i in slice m in the t-th observation window can be calculated as

$$r_{c_{m,i}}\left(t\right) = \sum_{\mathbf{s}_{m,j}(t)\in c_{m,i}(t)} n_{m,j}(t)\cdot d_{m,j}(t) \tag{6}$$

LSTM is an improved variation of recurrent neural network (RNN), which uses memory cells to solve the vanishing and exploding gradient problems. It consists of three components, i.e., the input gate $\mathbf{i_t}$ determining the extent of information to be written into the cell, the forget gate $\mathbf{f_t}$ determining the extent to forget the previous data, and the output gate $\mathbf{o_t}$ determining what output to generate from the current cell, and more details about LSTM can be found in [10]. Fig. 1 illustrates the structure of the LSTM cell memory used in this letter, where $\mathbf{x}_t$, $\mathbf{h}_t$, and $\mathbf{c}_t$ are the input, the output, and the cell state, respectively. In our problem, the output $\mathbf{h}_t$ of LSTM is a vector whose elements stand for the predicted traffic loads for user clusters in the $(t+1)$-th prediction window, i.e.,

$$\mathbf{h}_t = \left[\widehat{r}_{c_{1,1}}\left(t+1\right), ..., \widehat{r}_{c_{N,G}}\left(t+1\right)\right]^T \tag{7}$$

where $(\cdot)^T$ is the matrix transpose operation, and G is the total number of clusters of all NSs, i.e., $G = \sum_{m=1}^{M} K_m$. Meanwhile,

This article has been accepted for publication in IEEE Wireless Communications Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LWC.2022.3195517

the input is a vector $\mathbf{x}_t$ whose elements are the real measured traffic loads of all clusters in the previous N observation windows, which is given below

$$\mathbf{x}_t = \left[\begin{array}{l} r_{c_{1,1}}(t-N), \cdots, r_{c_{1,1}}(t) \\ \vdots \\ r_{c_{N,G}}(t-N), \cdots, r_{c_{N,G}}(t) \end{array} \right]. \tag{8}$$

In LSTM, the cell state $\mathbf{c}_t$ is determined by the forget gate $\mathbf{f}_t$ and the input gate $\mathbf{i}_t$, of which the expressions are respectively given as

$$\mathbf{f}_t = \sigma\left(\mathbf{W}_{fx}\mathbf{x}_t + \mathbf{W}_{fh}\mathbf{h}_{t-1} + \mathbf{b}_f\right) \tag{9}$$

and

$$\mathbf{i}_t = \sigma\left(\mathbf{W}_{ix}\mathbf{x}_t + \mathbf{W}_{ih}\mathbf{h}_{t-1} + \mathbf{b}_i\right) \tag{10}$$

where $\sigma(\cdot)$ is the sigmoid function, whose derivative is still a function of itself as $\sigma(z)' = \sigma(z)(1-\sigma(z))$. $\mathbf{W}$ and $\mathbf{b}$ are the weight matrix and bias vector, respectively. Then, we can get $\mathbf{c}_t$ as

$$\begin{array}{l} \tilde{\mathbf{c}}_t = \tanh\left(\mathbf{W}_{cx}\mathbf{x}_t + \mathbf{W}_{ch}\mathbf{h}_{t-1} + \mathbf{b}_c\right) \\ \mathbf{c}_t = \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \tilde{\mathbf{c}}_t \end{array} \tag{11}$$

where $\circ$ denotes the Hadamard (element-wise) product operation, and $\tanh(\cdot)$ is the hyperbolic tangent function, whose derivative is still a function of itself as $\tanh(z)' = 1 - \tanh(z)^2$.

Given the output gate expression as

$$\mathbf{o}_t = \sigma\left(\mathbf{W}_{ox}\mathbf{x}_t + \mathbf{W}_{oh}\mathbf{h}_{t-1} + \mathbf{b}_o\right). \tag{12}$$

we obtain

$$\mathbf{h}_t = \mathbf{o}_t \circ \tanh(\mathbf{c}_t). \tag{13}$$

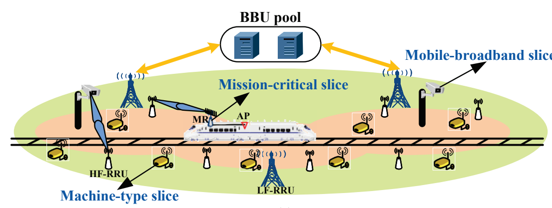
Suppose that the real measured traffic loads of user cluster $\mathbf{c}_{m,i}$ in the $(t+1)$-th observation window is $r_{c_{m,i}}(t+1)$, the loss function of the LSTM-based traffic load prediction model in this letter is defined as

$$L_t = \frac{1}{\sum\limits_{m=1}^{M} K_m} \sum_{m=1}^{M} \sum_{i=1}^{K_m} \left(r_{c_{m,i}}(t+1) - \widehat{r}_{c_{m,i}}(t+1) \right)^2. \tag{14}$$

III. Service-tailored Resource Reservation Mechanism

From [11], the management of virtualized radio resources in network slicing is at a MAC-frame granularity. Besides, the network usually can acquire the channel state information (CSI) of users, and hence can estimate the achieved capacity of radio resources themselves. Therefore, in this letter, we view the amount of radio resources and the transmission capacity the same for analysis purpose. Then, according to the traffic load prediction, we design a service-tailored resource reservation scheme for HSR wireless networks.

To support the dramatically growing mobile applications in HSR, the broadband millimeter wave (mmWave) technology will be employed in HSR wireless networks to boost the transmission capacity for some capacity demanding application scenarios, such as at HSR hub stations with video surveillance. Unfortunately, the directional radiation of mmWave communications will reduce the coverage and mobility performance of wireless networks especially under high-mobility HSR. In our previous study [12], [13], to solve this problem, we have proposed a sub-6GHz and mmWave bands integrated HSR wireless access network architecture. As shown in Fig. 2(a), the critical data, such as URLLC services, are transmitted by sub-6GHz bands with omni-directional coverage to guarantee the transmission reliability, while the large-volume data, such as eMBB services, are transmitted by mmWave bands with directional beams to enhance the transmission capacity. The whole network is built on the cloud radio access network (C-RAN) technology. The baseband resources are gathered in the BBU pool, and two kinds of remote radio units (RRUs), i.e., low-frequency RRUs (LF-RRUs) operating at sub-6GHz and high-frequency RRUs (HF-RRUs) operating at mmWave bands, are deployed along rails. Subsequently, we design the service-tailored resource reservation mechanism under this network architecture.



(a)

Type	HSR services	Mobility level	Preferred band
URLLC (Mission-critical slice)	Train-to-ground control data	High (100-138m/s)	Sub-6GHz
	Maintenance data of inspection vehicle	Medium (10-30m/s)	mmWave
	Dispatching data of ground workers	Low (1-3m/s)	Sub-6GHz
eMBB (Mobile-broadband slice)	Train-to-ground passenger data	High (100-138m/s)	mmWave
	Ground passenger data	Low (1-3m/s)	mmWave
	Wayside video surveillance data	Low (static)	mmWave
mMTC (Machine-type slice)	Train-to-ground sensing data	High (100-138m/s)	Sub-6GHz
	Inspection vehicle sensing data	Medium (10-30m/s)	mmWave
	Ground sensing data	Low (static)	Sub-6GHz

(b)

Fig. 2. Service-tailored resource reservation, (a) the sub-6GHz and mmWave bands integrated HSR wireless network, (b) Service categories and preferred resources.

In the two-stage traffic load prediction, the velocity feature of users can be extracted from the output mean vectors of the K-means based traffic flow clustering algorithm. Then, we can use it as a key reference for resource reservation to achieve a better match between users under different mobility and radio resources. For clarity, in Fig. 2(b), several HSR services are taken as case study to show which band resources should be reserved for which services under different mobility levels. In the sub-6GHz and mmWave bands integrated HSR wireless network, sub-6GHz bands can provide omni-directional coverage, and hence are more suitable to serve mission-critical services under high mobility. Consequently, according to the velocity information from different clusters, the mission-critical services with high mobility, such as train-to-ground control data and sensing data, are assigned with sub-6GHz resources. After the resource reservation for high-mobility URLLC service, if there are still sub-6GHz resources left, they will be reserved for other URLLC users (such as the dispatching data of ground workers) to better guarantee the transmission reliability or for low-data-rate mMTC users to save energy (such as ground sensing data). In contrast, broadband mmWave resources should be employed to carry the services with less-stringent reliability requirements but high capacity requirements, such as train-to-ground passenger data, high-volume wayside video surveillance data, etc. For other HSR services, the same resource reservation method can also be applied. From this design, we observe that the traffic flow clustering and the traffic load prediction provide the fundamental and valuable references for the NS resource reservation, improving radio resource utilization while guaranteeing the transmission reliability for sliced HSR wireless networks.

IV. Numerical Results

In this section, we present numerical results and the corresponding performance analysis for our proposed two-stage traffic load prediction scheme. Without loss of generality, in each of the three NSs, we set three kinds of traffic patterns, and the corresponding parameter values of these traffic flows are listed in Table I. The observation window T_0 is 10min [6], the hidden units of LSTM is 100, and the learning rate of LSTM is 0.1. A total of 1600 observation window data are generated as the data training sets of the K-means traffic flow clustering algorithm. After clustering, the traffic load data calculated through Eq.(6) are subsequently taken as the input to train the LSTM model. Since our main idea is to cluster traffic flows before prediction, for fairness, we compare our proposed scheme with the conventional LSTM scheme, namely LSTM only scheme, where the traffic load data of different NSs are directly input into the LSTM algorithm for prediction without clustering.

Since our proposed scheme involves two stages, i.e., traffic flow clustering and traffic load prediction, we present the simulation results to respectively show the performance of the two stages. In the traffic flow clustering stage, we use the silhouette coefficient to evaluate the clustering validity [14], and then determine the optimal value of K. For any sample $\mathbf{s}_{m,i}(t)$, its silhouette coefficient is calculated as $sil(i) = (b(i) - a(i))/(max(a(i), b(i)))$, where $a(i)$ denotes the average distance of $\mathbf{s}_{m,i}(t)$ to the other samples within the same cluster, while $b(i)$ denotes the average distance of $\mathbf{s}_{m,i}(t)$ to the other samples within its closest neighboring cluster. $sil(i)$ ranges from -1 to 1. When $a(i)$ is much smaller than $b(i)$, which means the distance of the data point to its own cluster is much smaller than that to other clusters, $sil(i)$ is close to 1 to show this data point is well clustered. Oppositely, $sil(i)$ is close to -1 to show it is badly clustered. Then, the silhouette value of the whole cluster is defined as $S(\mathbf{D}_m) = \frac{1}{U_m} \sum_{\mathbf{s}_{m,i} \in \mathbf{D}_m} sil(i)$. As a case study, Fig. 3 depicts the clustering silhouette value of the URLLC service. In the simulation, we set three kinds of traffic patterns for URLLC. From the results in Fig. 3, we observe that the optimal value of K is also 3, implying that the K-means algorithm can properly cluster the URLLC services. In the following simulations, we set $K_m = 3$.

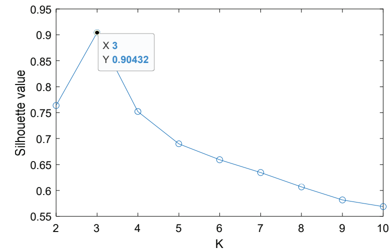


Fig. 3. The clustering silhouette value of the URLLC service.

Next, in Fig. 4, the traffic load and average velocity of URLLC services in the obtained clustering mean vectors are shown. Through the clustering results, we acquire the mobility status of different user clusters. As discussed in Section III, based on the mobility status of users, we can reserve proper network resources for them to improve their mobility performance. For instance, sub-6GHz resources are preferentially reserved for high-mobility URLLC users, while mmWave resources are suitable for low-mobility eMBB users.

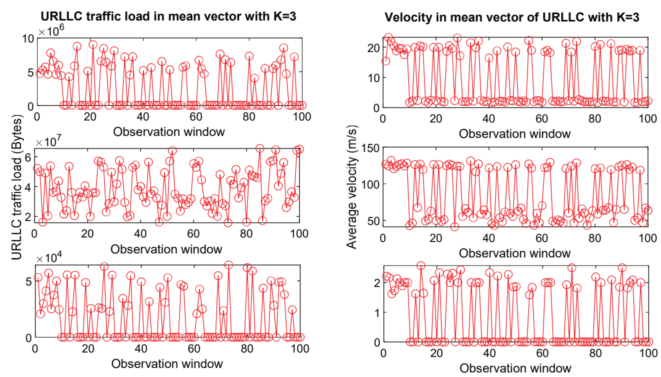


Fig. 4. The clustering mean vector of URLLC service.

Then, in Fig. 5, we compare the traffic load prediction performance of our proposed scheme with the conventional LSTM only scheme, where Figs. 5(a)-(d) present the traffic load prediction results of URLLC service under LSTM only scheme, the traffic load prediction results of URLLC service under the proposed scheme, the traffic load prediction results of all three kinds of services under LSTM only scheme, and the traffic load prediction results of all three kinds of services under the proposed scheme, respectively. For clarity, only the results in the first 100 observation windows are depicted. From the results, we observe that owing to the pre-processing of the traffic flow clustering in the proposed scheme, the LSTM algorithm can better capture the variations of traffic loads and achieve higher prediction accuracy, implying higher NS resource reservation accuracy.

Finally, in Fig. 6, we illustrate the overall traffic load prediction accuracy of two different schemes. From the results, we observe that compared with the LSTM only scheme, our proposed scheme can significantly improve the prediction accuracy. Therefore, the conclusion can be drawn that the pre-processing of traffic flow clustering can significantly improve the prediction accuracy, therefore enhancing the NS resource reservation performance.

TABLE I
SIMULATION PARAMETERS

Parameter	Value
Mission-critical slice	$n(t)$: uniform distributions $[4000, 4500]/T_0$, $[1500, 2000]/T_0$, $[500, 600]/T_0$ $d(t)$: uniform distributions [500,600] bytes, [100,200] bytes, [20,30] bytes $v(t)$: uniform distributions [100,138] m/s, [10,30] m/s, [1,3] m/s $U(t)$: uniform distributions [2,5] users, [2,8] users, [6,16] users
Mobile-broadband slice	$n(t)$: Poisson distributions $6000/T_0$, $2400/T_0$, $240/T_0$ $d(t)$: uniform distributions [3000,3500] bytes, [300,400] bytes, [100,200] bytes $v(t)$: uniform distributions [100,138] m/s, [1,3] m/s, 0 $U(t)$: uniform distributions [2,5] users, [5,10] users, [8,20] users
Machine-type slice	$n(t)$: uniform distributions $[200, 500]/T_0$, $[1000, 1100]/T_0$, $[10, 100]/T_0$ $d(t)$: uniform distributions [100,110] bytes, [200,210] bytes, [50,70] bytes $v(t)$: uniform distributions [100,138] m/s, [10,30] m/s, 0 $U(t)$: uniform distributions [4,10] users, [1,10] users, [1,20] users

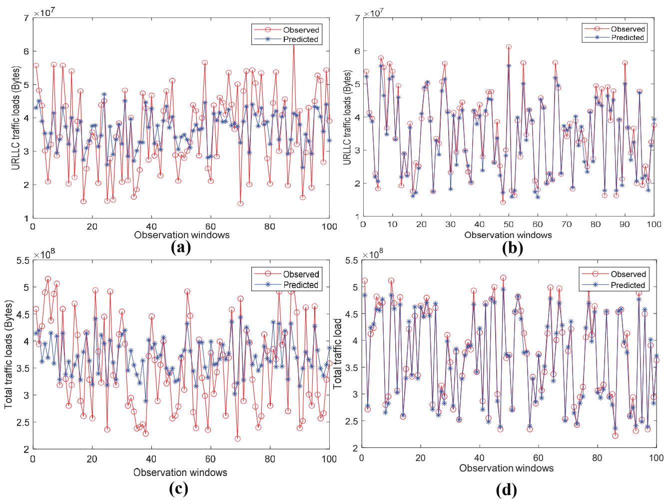


Fig. 5. Traffic load prediction results, (a) URLLC under LSTM only scheme, (b) URLLC under the proposed scheme, (c) all three services under LSTM only scheme, (d) all three services under the proposed scheme.

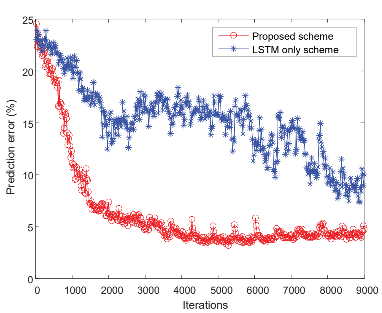


Fig. 6. Prediction accuracy comparison.

V. CONCLUSIONS AND PROSPECTIVES

In this letter, we propose a two-stage traffic load prediction scheme, where in the first stage, the K-means algorithm is leveraged to perform intra-NS user clustering, while in the second stage, the LSTM algorithm is applied to predict traffic loads. Then, based on the clustering and prediction results, we design a service-tailored resource reservation mechanism for the sub-6GHz and mmWave bands integrated HSR wireless networks. Our numerical results have demonstrated that our proposed scheme can significantly improve traffic load prediction accuracy to ensure the NS resource reservation performance. In our future work, based on the results of this letter, we will study the adaptive and intelligent resource allocation for sliced HSR wireless networks.

REFERENCES

[1] B. Ai, A. F. Molisch, M. Rupp, and Z.-D. Zhong, "5G key technologies for smart railways," *Proceedings of the IEEE*, vol. 108, no. 6, pp. 856–893, 2020.
[2] J. Lee, J. Park, J. Kim, and J. Lee, "Consideration on physical layer aspects of NR systems in HST scenarios," in *IEEE 90th Vehicular Technology Conference (VTC2019-Fall), Honolulu, HI, USA*, 2019, pp. 1–5.
[3] E. Berlin and K. Van Laerhoven, "Sensor networks for railway monitoring: Detecting trains from their distributed vibration footprints," in *IEEE International Conference on Distributed Computing in Sensor Systems, Cambridge, MA, USA*, 2013, pp. 80–87.
[4] G.S. Kuo, P.C. Ko, and M.L. Kuo, "A probabilistic resource estimation and semi-reservation scheme for flow oriented multimedia wireless networks," *IEEE Communications Magazine*, vol. 39, no. 2, pp. 135–141, 2001.
[5] S. Cho, Y. Kim, and Y. Han, "Analysis of control overhead reduction by semi-persistent scheduling," *IEEE Wireless Communications Letters*, vol. 9, no. 5, pp. 726–730, 2020.
[6] H. Zhang and V. W. S. Wong, "A two-timescale approach for network slicing in C-RAN," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 6, pp. 6656–6669, 2020.
[7] K. Zhu and E. Hossain, "Virtualization of 5G cellular networks as a hierarchical combinatorial auction," *IEEE Transactions on Mobile Computing*, vol. 15, no. 10, pp. 2640–2654, 2016.
[8] W. Wang, C. Zhou, H. He, W. Wu, W. Zhuang, and X. Shen, "Cellular traffic load prediction with LSTM and gaussian process regression," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC), Dublin, Ireland*, 2020, pp. 1–6.
[9] N. Shi, X. Liu, and Y. Guan, "Research on K-means clustering algorithm: An improved K-means clustering algorithm," in *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, 2010, pp. 63–67.
[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[11] R. Kokku, R. Mahindra, H. Zhang, and S. Rangarajan, "NVS: A substrate for virtualizing wireless resources in cellular networks," *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1333–1346, 2012.
[12] L. Yan, X. Fang, X. Wang, and B. Ai, "AI-enabled sub-6-GHz and mm-Wave hybrid communications: Considerations for use with future HSR wireless systems," *IEEE Vehicular Technology Magazine*, vol. 15, no. 3, pp. 59–67, 2020.
[13] L. Yan, X. Fang, L. Hao, and Y. Fang, "A fast beam alignment scheme for dual-band HSR wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 4, pp. 3968–3979, 2020.
[14] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," in *IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Sydney, NSW, Australia*, 2020, pp. 747–748.