Received 16 December 2021; revised 11 January 2022; accepted 21 January 2022. Date of publication 28 January 2022; date of current version 7 February 2022.

Digital Object Identifier 10.1109/OJCOMS.2022.3146886

Traffic Prediction-Enabled Energy-Efficient Dynamic Computing Resource Allocation in CRAN Based on Deep Learning

YONGQIN FU (Graduate Student Member, IEEE), AND XIANBIN WANG (Fellow, IEEE)

Department of Electrical and Computer Engineering, Western University, London, ON N6A 5B9, Canada

CORRESPONDING AUTHOR: X. WANG (e-mail: xianbin.wang@uwo.ca)

This work was supported in part by the Discovery Program of Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant RGPIN2018-06254, and in part by the Canada Research Chair Program.

ABSTRACT Due to the greatly increased bandwidth of 5G networks compared with that of 4G networks, the power consumption brought by baseband signal processing of 5G networks is much higher, which inevitably raises the operation expenditures. Cloud Radio Access Network (CRAN) is widely adopted in 5G networks, which splits the traditional base stations into Remote Radio Heads (RRHs) and Baseband Units (BBUs), which are equipped with computing resource for baseband signal processing. The number of required BBUs varies due to the fluctuation of wireless traffic of RRHs. Hence, fixed computing resource allocation might waste power. This paper investigates energy-efficient dynamic computing resource allocation in CRAN by predicting the wireless traffic of RRHs and allocating computing resource based on the prediction results aiming at using fewest BBUs to minimize power consumption. For wireless traffic prediction, a novel method based on two-dimensional CNN LSTM model with temporal aggregation is proposed. By treating the wireless traffic data as images, this model could extract spatial correlation from these data to improve accuracy. Moreover, the problem of dynamic computing resource allocation in CRAN is formulated as an offline four-constraint bin packing problem, considering both uplink and downlink baseband signal processing capacities of BBUs and Common Public Radio Interface (CPRI) bandwidths. For solving this problem, a Multi-start Simulated Annealing (MSA) algorithm is proposed. Simulation results demonstrate that the proposed method for wireless traffic prediction could outperform the state-of-the-art deep learning models. In addition, the proposed MSA algorithm could achieve lower power consumption than the state-of-the-art heuristic algorithms.

INDEX TERMS Computing resource allocation, wireless traffic prediction, CRAN, deep learning, two-dimensional CNN LSTM, multi-start simulated annealing.

I. INTRODUCTION

POWER consumption accounts for an important part of the expenditures of the network operators. According to [1], the network operation expenditure takes up about 25% of the total cost base of the network operators, over 90% of which is spent on power consumption. Moreover, 82% - 97% of the power consumption in the network is spent on powering the Radio Access Network (RAN) [1], where baseband signal processing is conducted. The situation even gets worse in the 5G era. The power consumption of a typical 5G base station is up to twice or more of that of a 4G base station [2]. Compared with that of 4G networks, the bandwidth of 5G networks have been greatly increased, which makes the power consumption brought by baseband signal processing in 5G networks much higher than that of 4G networks. Hence, reducing the power consumption brought by baseband signal processing at RAN could help to reduce the network operation expenditures of the network operators to prompt the commercial application of 5G technologies, as well as mitigate climate change.

The power consumption of baseband signal processing is influenced by the architecture of RAN. Cloud Radio Access Network (CRAN) [3], [4] is an architecture of radio access network widely adopted in 5G networks, which is illustrated

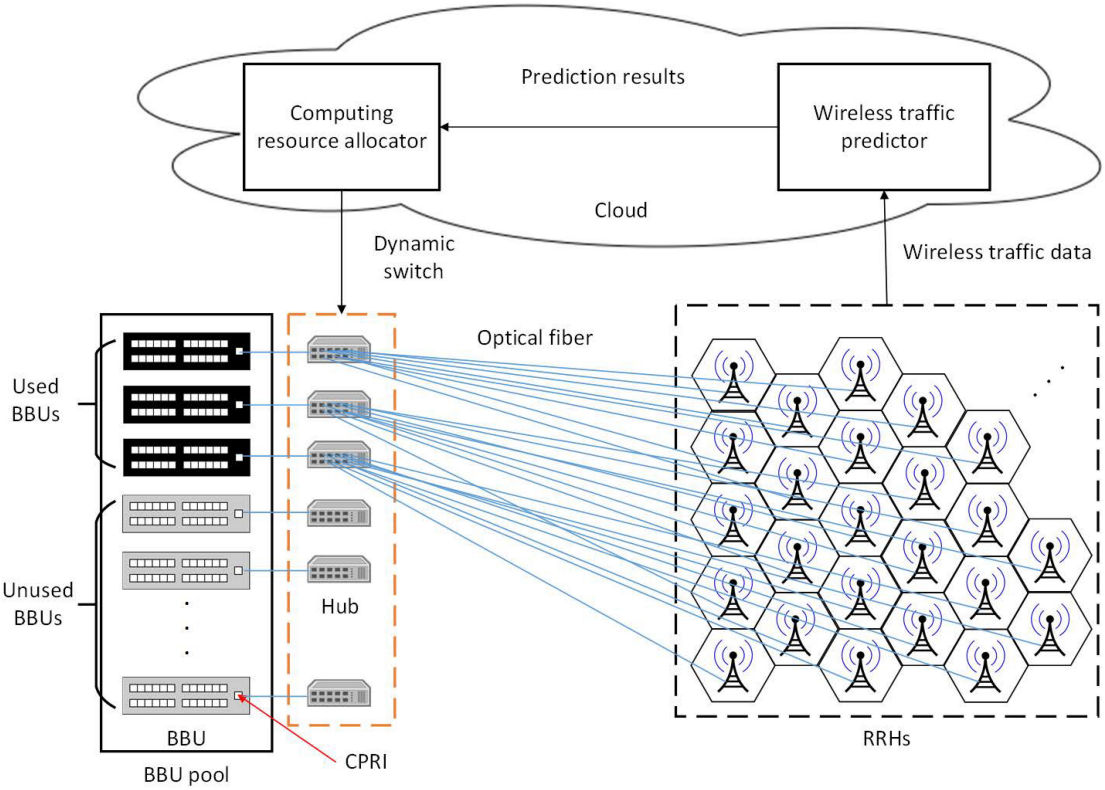


FIGURE 1. Illustration of energy-efficient dynamic computing resource allocation in CRAN enabled by wireless traffic prediction.

in Fig. 1. As can be seen in Fig. 1, in CRAN, the traditional base stations are split into two parts, which are Remote Radio Heads (RRHs) and Baseband Units (BBUs). RRHs are responsible for transmitting and receiving the wireless signals, while BBUs are responsible for baseband signal processing. Multiple BBUs form a BBU pool, which makes it easier for dynamic allocation of computing resource. Moreover, by being connected to the same hub which is connected to one BBU, multiple RRHs can be served by one BBU simultaneously for baseband signal processing. When a BBU is not in use during one time period, it could be temporarily switched off in order to reduce power consumption.

Apparently, the more RRHs one BBU serves simultaneously, the less amount of BBUs are used and the less power the BBU pool consumes. However, there exist four constraints which limit the minimum number of BBUs required to serve all the RRHs, which are uplink and downlink baseband signal processing capacities of BBUs and Common Public Radio Interface (CPRI) bandwidths. For the first two constraints, the amounts of computing resource required for uplink and downlink baseband signal processing of one RRH are proportional to its uplink and downlink wireless traffic, respectively. In this paper, the computing resource in BBUs is measured in Giga Operations Per Second (GOPS). For the last two constraints, the amounts of uplink and downlink CPRI bandwidths consumed by baseband signal processing of one RRH are also proportional to its uplink and downlink wireless traffic, respectively. As illustrated in Fig. 1, the RRHs are connected to hubs through optical fibers and each hub is connected to one BBU through Common Public Radio Interface (CPRI). Because RRHs don't compress the signals, high-speed In-phase/Quadrature (I/Q) sampling signals need to be transmitted between RRHs and BBUs, which requires high bandwidth. The uplink and downlink CPRI bandwidths determine the maximum amounts of I/Q sampling signals which can be transmitted to and from BBU through CPRI simultaneously, respectively.

Hence, in order to minimize the total number of used BBUs to reduce power consumption, the wireless traffic of RRHs needs to be accurately estimated. Moreover, the wireless traffic of RRHs doesn't stay unchanged. On the contrary, in real-world scenarios, the wireless traffic of RRHs fluctuates greatly during one day. For example, the wireless traffic of RRHs at peak time is high, but the wireless traffic of RRHs at off-peak time is low.

Therefore, prediction of the wireless traffic of the RRHs in advance is required to facilitate the process of computing resource allocation, aiming at minimizing the total number of used BBUs to minimize the total power consumption of the BBU pool. Wireless traffic prediction is an important engineering problem, which is in fact a time series prediction problem. Many techniques have been developed by both the academia and industry for solving the time series prediction problem. Traditional time series prediction techniques include classic models, machine

learning techniques and deep learning techniques. Classic models include Autoregressive Moving Average (ARMA) models [5], Autoregressive Integrated Moving Average (ARIMA) models [6] and Seasonal Autoregressive Integrated Moving Average (SARIMA) models [7]. Machine learning techniques for time series prediction include Support Vector Machine (SVM) [8], K Nearest Neighbor (KNN) Regression [9] and Classification and Regression Trees (CART) [10]. Deep learning techniques for time series prediction include Recurrent Neural Network (RNN) [11], Long Short-Term Memory (LSTM) [12], [13], Gated Recurrent Unit (GRU) [14], deep autoencoder [15], Restricted Boltzmann Machine (RBM) [16] and Deep Belief Network (DBN) [17]. Except for these, temporal aggregation [18] is another useful technique for time series prediction, which has the potential to improve the prediction accuracy. Utilizing temporal aggregation, the prediction of lower-frequency series could be achieved by predicting the higher-frequency series and aggregating their prediction results.

Once the wireless traffic prediction process is completed, the demands on uplink and downlink computing resources and CPRI bandwidths of RRHs could be estimated, which facilitates the process of computing resource allocation. Since the aim of dynamic computing resource allocation is to minimize the total number of used BBUs in order to minimize the total power consumption of the BBU pool, this problem could be formulated as an offline four-constraint bin packing problem, which is NP-hard. For solving the bin packing problem, there exist three kinds of techniques, which are approximation algorithms [19], [20], metaheuristic algorithms and deep reinforcement learning [21]. Approximation algorithms include Next Fit (NF), Worst Fit (WF), First Fit (FF), Best Fit (BF), First Fit Decreasing (FFD) and Best Fit Decreasing (BFD). Metaheuristic algorithms include Simulated Annealing (SA) [22], Genetic Algorithm (GA) [23] and Particle Swarm Optimization (PSO) [24].

The main contributions of this paper can be summarized as follows:

1) In order to better reflect the real-world scenarios, the problem of energy-efficient dynamic computing resource allocation in CRAN is formulated as an offline four-constraint bin packing problem, in which the uplink and downlink baseband signal processing capacities of BBUs and CPRI bandwidths are regarded as constraints. Moreover, the differences among the demands on computing resource and CPRI bandwidth of different types of wireless traffic are distinguished.
2) A method for wireless traffic prediction is proposed based on two-dimensional CNN LSTM model with temporal aggregation. By treating the wireless traffic data as images, this model could extract the spatial correlation among the wireless traffic of adjacent RRHs. Moreover, by employing temporal aggregation, this model could capture the relationship among the wireless traffic during the time slots in the current time period and the wireless traffic during those in the upcoming time period to improve the prediction accuracy. In addition, by using the sum of the prediction results of wireless traffic during the time slots in the upcoming time period as the prediction result, their prediction errors could be offset, which further improves the prediction accuracy.
3) A multi-start simulated annealing (MSA) algorithm is proposed for solving the formulated offline four-constraint bin packing problem. By randomly selecting multiple starts, the MSA algorithm could mitigate the influence of the position of start on the final result to reduce the number of used BBUs in order to reduce power consumption of the BBU pool.
4) To conduct performance evaluation more realistically and accurately, a real-world dataset is utilized for validating the effectiveness of our proposed wireless traffic prediction method and MSA algorithm. The numerical results confirm that our proposed wireless traffic prediction method based on two-dimensional CNN LSTM model with temporal aggregation could achieve more accurate prediction performance compared with the state-of-the-art deep learning models. Moreover, the numerical results confirm that our proposed MSA algorithm could achieve lower power consumption of the BBU pool compared with the state-of-the-art metaheuristic algorithms.

II. RELATED WORK

In this section, we will introduce the recent progress of research in wireless traffic prediction based on deep learning and computing resource allocation in CRAN.

A. WIRELESS TRAFFIC PREDICTION BASED ON DEEP LEARNING

Due to the promising potential of deep learning on wireless traffic prediction, researchers from both the academia and the industry have devoted a lot of effort to proposing deep learning models for wireless traffic prediction. A novel deep learning architecture named DenseNet was proposed in [25], which can effectively capture the complex patterns hidden in cellular data. A Spatial Temporal neural Network (STN) is proposed in [26], which is a precise cellular traffic forecasting architecture. And a Double STN (D-STN) is proposed, which employs a light-weight mechanism for combining the STN output with historical statistics, thereby improving long-term prediction performance.

Although most proposed deep learning models for wireless traffic prediction are based on variants of LSTM model, some other deep learning techniques have also been applied for wireless traffic prediction. A novel deep neural network architecture, STCNet, is proposed in [27] for cellular traffic prediction, which contains ConvLSTM units to simultaneously capture the spatial and temporal dependencies of cellular traffic. Moreover, transfer learning strategy is also utilized for exploiting the similarities among different

types of cellular traffic as well as capturing the pattern similarity of cellular traffic among different areas. A meta-learning scheme is proposed in [28], aiming at addressing the problem of short-term user-level network traffic prediction. The proposed meta-learning scheme is an ensemble of different predictors for predicting different types of traffic. Moreover, deep reinforcement learning is utilized for choosing the appropriate predictor according to the recent prediction performance. The authors in [29] propose a cross-service and regional fusion transfer learning strategy in order to utilize multiple cross-domain datasets for enhancing the prediction performance. The authors in [30] propose a spatial-temporal deep learning method for citywide cellular traffic prediction, which incorporates the attention scheme in the model architecture design. Taking the effect of handover into consideration, in [31], a novel cellular traffic prediction model, STGCN-HO, is proposed, utilizing the handover graph together with a stacked residual neural network structure in order to improve the prediction performance. The authors in [32] propose a novel framework for wireless traffic prediction, called FedDA, which employs federated learning to train the wireless traffic prediction model in a collaborative way. Moreover, a dual attention scheme is proposed for constructing the global model through the aggregation of intra-cluster and inter-cluster models. In [33], the authors utilize 5 different types of deep learning models for wireless traffic usage forecasting, which are LSTM, GRU, CNN, CNN-LSTM and CNN-GRU. The forecasting results demonstrate that the last two models where CNN is utilized for extracting spatial dependencies existing among neighboring access points (APs) could forecast the wireless traffic usage of a single AP when significant spatial correlations exist.

Some researchers combine statistical tools with deep learning models for improving prediction accuracy. In [34], a novel single-cell level cellular traffic prediction method is proposed by combining LSTM with Gaussian Process Regression (GPR) for improving the prediction performance. In this method, the dominant periodic components are extracted utilizing Fourier analysis. In addition, LSTM is utilized for learning the long-term dependency among the small random values and GPR is applied for estimating the residual random components. A novel wireless traffic prediction architecture is proposed in [35], aiming at improving prediction accuracy by series fluctuation pattern clustering. Moreover, a novel model based on LSTM, called TPBLN, is proposed for predicting the baseline component, while the residual component is predicted through a probability model, whose parameter is estimated by adopting the maximum likelihood estimation method.

B. COMPUTING RESOURCE ALLOCATION IN CRAN

Computing resource allocation in CRAN has attracted a lot of attention from both the academia and industry. In [36], a novel BBU-RRH association scheme is proposed for minimizing the power consumption. Graph partitioning and rejoining is utilized in the proposed scheme, which could reduce the communication overhead for energy saving. The authors in [37] investigate the problem of computational resources allocation between RRHs and BBUs with the aim of reducing the number of used BBUs and maximizing the data rates. They decompose this problem into two sub-problems, which are about resource block allocation and BBU allocation, respectively. For BBU allocation, the first fit decreasing algorithm is utilized for minimizing the number of used BBUs. The authors in [38] formulate the BBU processing allocation problem as a bin packing problem, aiming at minimizing the total power consumption with guarantee of per-user QoS. For solving this problem, a two-phase heuristic algorithm is proposed based on BFD algorithm. In [39], to maximize the computing resource utilization, computing resource allocation among BBUs is formulated as a game-theory bargaining problem, taking the Quality of Service (QoS) demands of services into consideration, which is solved by a generalized Nash bargaining solution.

Some researchers utilize wireless traffic prediction for computing resource allocation in CRAN. The authors in [40] investigate the problem of allocating multiple BBU pools to multiple RRHs, aiming at raising resource utilization and reducing power consumption. LSTM is utilized for predicting the throughput of each RRH and a GA-based resource allocation algorithm (GARAA) is proposed for the allocation of BBU pools to minimize power consumption. Aiming at maximizing the QoS and minimizing the blocked connections, the authors in [41] formulate the problem of dynamic BBU-RRH mapping and use a Markov model for cell load prediction and a genetic algorithm for solving the problem based on the prediction results.

Some researchers take load balancing among BBUs into consideration. In [42], the problem of dynamic BBU-RRH mapping is formulated as an optimization problem, aiming at achieving load balancing among BBUs and improving the QoS. This problem is solved by utilizing Genetic Algorithm (GA). The authors in [43] investigate the problem of BBU computing resource allocation aiming at minimizing the total amount of used computing resources and concurrently balancing the use of computing resources among all the used BBUs. To solve the problem formulated equivalent to the classical bin-packing problem, they propose a heuristic genetic algorithm (HeuGA), which combines the FF algorithm with GA algorithm. In [44], the authors investigate the problem of dynamic BBU-RRH mapping and formulate it as a linear integer-based constrained optimization problem with the aim of achieving load balancing among BBUs and avoiding unnecessary handovers, which is solved by an Estimation Distribution Discrete Particle Swarm Optimization (EDDPSO) algorithm.

Some researchers also consider to turn off idle RRHs for energy saving. In [45], the problem of real-time BBU/RRH assignment for CRAN in LTE is investigated, aiming at minimizing the number of active BBUs in order to save energy. This problem is formulated as a multiple knapsack problem and solved by utilizing IBM's linear solver CPLEX.

III. SYSTEM MODEL AND PROBLEM FORMULATION

A. SYSTEM MODEL

In this paper, we consider an area which is divided into M cells. For each cell, one RRH is located at the center. All the RRHs are served by one BBU pool, which is consisted of N BBUs. As illustrated in Fig. 1, the RRHs are connected to hubs through optical fibers and each hub is connected to one BBU through CPRI. The connections from RRHs to hubs can be flexibly switched according to decisions of computing resource allocator. Moreover, we consider both the uplink and downlink wireless traffic of three different types of services, which are calls, Short Messaging Service (SMS) and Internet.

The total power consumption of the BBU pool during time period t is:

$$P_{total}(t) = \sum_{i=1}^{N} P_i(t) + P_{pool}^{basic}(t), \tag{1}$$

where $P_{pool}^{basic}(t)$ is the basic power consumption required for operating the BBU pool during time period t and $P_i(t)$ denotes the power consumption of the i-th BBU in the BBU pool during time period t, which is:

$$P_i(t) = u_i(t)\left(\sum_{k=1}^{6} \lambda_k s_{i,k}(t) + P_{BBU}^{basic}(t)\right), \tag{2}$$

where $u_i(t) = 1$ or 0 indicates the i-th BBU is in use or not in use during time period t, respectively. Moreover, $s_{i,k}(t)$ denotes the volume of the k-th type of wireless traffic of the RRHs served by the i-th BBU during time period t. In addition, λ_k is a coefficient for determining the power consumption introduced by baseband signal processing of the k-th type of wireless traffic and $k = 1, 2, 3, 4, 5, 6$ correspond to uplink call, downlink call, uplink SMS, downlink SMS, uplink Internet traffic and downlink Internet traffic, respectively. Besides, $P_{BBU}^{basic}(t)$ is the basic power consumption required for operating a single BBU during time period t.

The volume of the k-th type of wireless traffic processed by the i-th BBU during time period t is:

$$s_{i,k}(t) = \sum_{j=1}^{M} C_{i,j}(t) s_{j,k}(t), \tag{3}$$

where $C_{i,j}(t) = 1$ or 0 indicates the j-th RRH is served by the i-th BBU or not during time period t, respectively. Moreover, $s_{j,k}(t)$ denotes the volume of the k-th type of wireless traffic of the j-th RRH during time period t.

There exist constraints of the largest total amounts of uplink and downlink wireless traffic of RRHs which could be processed by a single BBU. The computing resource in BBUs is measured in Giga Operations Per Second (GOPS). Moreover, we assume that the maximum uplink and downlink baseband processing capacities of a single BBU are R_{up}^{max} and R_{down}^{max} GOPS, respectively. The uplink and downlink requirements on computing resource of the j-th RRH are:

$$L_j^{up}(t) = \alpha_1 s_{j,1}(t) + \alpha_3 s_{j,3}(t) + \alpha_5 s_{j,5}(t), \tag{4}$$

and

$$L_j^{down}(t) = \alpha_2 s_{j,2}(t) + \alpha_4 s_{j,4}(t) + \alpha_6 s_{j,6}(t), \tag{5}$$

respectively, where α_k is a coefficient for determining the amount of computing resource required for baseband signal processing of the k-th type of wireless traffic.

As mentioned before, the BBUs are connected to hubs through CPRIs. There exist constraints on both the maximum uplink and downlink bandwidths of one CPRI. In this paper, we assume that the maximum uplink and downlink bandwidths of one CPRI are W_{up}^{max} and W_{down}^{max} Gbps, respectively. The uplink and downlink CPRI bandwidth requirements of the j-th RRH are:

$$B_j^{up}(t) = \beta_1 s_{j,1}(t) + \beta_3 s_{j,3}(t) + \beta_5 s_{j,5}(t), \tag{6}$$

and

$$B_j^{down}(t) = \beta_2 s_{j,2}(t) + \beta_4 s_{j,4}(t) + \beta_6 s_{j,6}(t), \tag{7}$$

respectively, where β_k is a coefficient for determining the amount of CPRI bandwidths required for baseband signal processing of the k-th type of wireless traffic.

B. PROBLEM FORMULATION

Aiming at minimizing the total power consumption of the BBU pool, the problem of energy-efficient dynamic computing resource allocation in CRAN could be formulated as shown below.

$$\begin{aligned}
\text{Minimize} \quad & P_{total}(t) \\
\text{Subject to:} \quad & C1: u_i(t) \in \{0,1\}, \quad \forall i \\
& C2: C_{i,j}(t) \in \{0,1\}, \quad \forall i,j \\
& C3: \sum_{i=1}^{N} C_{i,j}(t) u_i(t) = 1, \quad \forall j \\
& C4: \sum_{j=1}^{M} C_{i,j}(t) L_j^{up}(t) \le R_{up}^{max}, \quad \forall i \\
& C5: \sum_{j=1}^{M} C_{i,j}(t) L_j^{down}(t) \le R_{down}^{max}, \quad \forall i \\
& C6: \sum_{j=1}^{M} C_{i,j}(t) B_j^{up}(t) \le W_{up}^{max}, \quad \forall i \\
& C7: \sum_{j=1}^{M} C_{i,j}(t) B_j^{down}(t) \le W_{down}^{max}, \quad \forall i.
\end{aligned}$$

In the problem formulation, the constraint C1 indicates that the indicator of the status of the i-th BBU should be either 0 or 1, meaning the i-th BBU is in use and not in use during time period t, respectively. The constraints C2 and C3 constrain that during time period t, each RRH

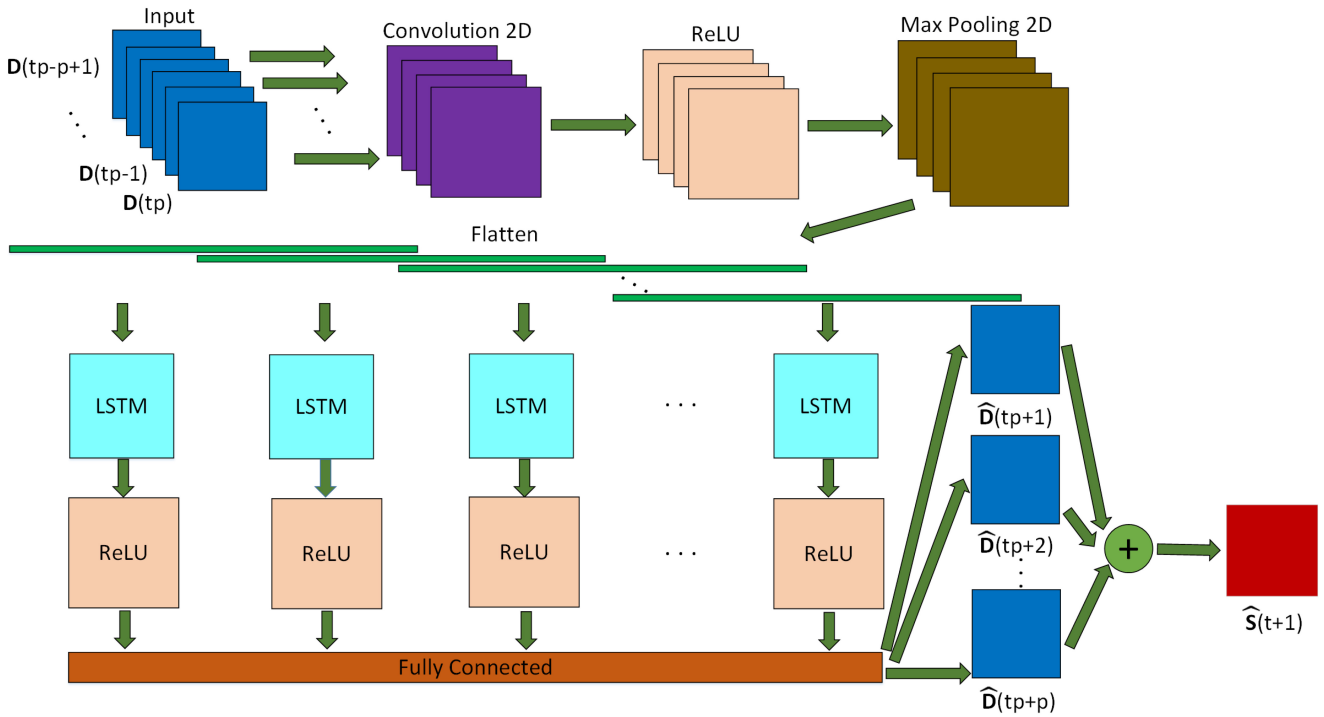


FIGURE 2. Illustration of the framework of the proposed wireless traffic prediction method.

should be and at most be served by one BBU which is in use. The constraints C4 and C5 constrain that for each BBU, the requirements of computing resource for uplink and downlink baseband signal processing of all the RRHs served by it cannot exceed its maximum uplink and downlink baseband signal processing capacities, respectively. The constraints C6 and C7 constrain that for each BBU, the requirements of uplink and downlink CPRI bandwidths of all the RRHs served by it cannot exceed the maximum uplink and downlink CPRI bandwidths, respectively.

IV. WIRELESS TRAFFIC PREDICTION BASED ON DEEP LEARNING

From Section III, we know that to successfully perform the task of energy-efficient dynamic computing resource allocation in CRAN, the wireless traffic of the RRHs needs to be predicted accurately. In this paper, we propose a novel method for wireless traffic prediction based on two-dimensional CNN LSTM model with temporal aggregation.

Long Short-Term Memory (LSTM) is a special type of Recurrent Neural Network (RNN), which has superiority over conventional RNN in long sequence prediction. LSTM has the ability of forgetting some unnecessary features and keep the necessary features for prediction. Moreover, the problems of gradient vanishment and gradient explosion are both taken into consideration in the design of the architecture of LSTM.

Convolutional Neural Network (CNN) is a classic deep learning model which has achieved great success in the field of computer vision. CNN utilizes convolutional operations for learning spatial relationships from images. Because there exist spatial correlation among the wireless traffic of RRHs located close to each other, the wireless traffic data of RRHs could be treated as images and CNN could be utilized for extracting spatial correlation from them, which are further processed by LSTM for wireless traffic prediction.

The framework of the proposed wireless traffic prediction method is illustrated in Fig 2. As illustrated in Fig. 2, the proposed deep learning prediction model is consisted of seven main layers. Moreover, in order to improve the prediction accuracy, temporal aggregation is utilized in the design of the proposed method. Specifically, one time period with length T is divided into p time slots with length $\frac{T}{p}$ and the wireless traffic during the past p time slots are utilized to predict the wireless traffic in the upcoming p time slots, which are aggregated together to generate the prediction result of the wireless traffic in the next time period. In Fig. 2, $\boldsymbol{D}(k)$ represents the wireless traffic matrix of a geographical area covered by $H \times W$ cells during the k-th time slot, which could be written as:

$$\boldsymbol{D}(k) = \begin{pmatrix} d_{1,1}(k) & \cdots & d_{1,W}(k) \\ \vdots & \ddots & \vdots \\ d_{H,1}(k) & \cdots & d_{H,W}(k) \end{pmatrix}, \tag{8}$$

where $d_{i,j}(k)$ denotes the wireless traffic of cell (i, j) during the k-th time slot. Moreover, $\boldsymbol{S}(t)$ represents the wireless traffic matrix during the t-th time period, which could be

written as:

$$\boldsymbol{S}(t) = \begin{pmatrix} s_{1,1}(t) & \cdots & s_{1,W}(t) \\ \vdots & \ddots & \vdots \\ s_{H,1}(t) & \cdots & s_{H,W}(t) \end{pmatrix}, \tag{9}$$

where $s_{i,j}(t)$ denotes the wireless traffic of cell (i, j) during the t-th time period.

The wireless traffic data of the target area during the past p time slots are used as inputs of the model, which are processed by a two-dimensional convolutional layer activated by the Rectified Linear Unit (ReLU) function. Next, a two-dimensional max pooling layer performs downsampling on the processed data, followed by a flatten layer which transforms the multi-dimensional input data into one-dimensional data. Then the generated one-dimensional data is processed by an LSTM layer activated by ReLU function, followed by a fully connected layer to generate the prediction results of the wireless traffic in the upcoming p time slots, which are added together to calculate the prediction result of wireless traffic in the next time period. By utilizing the wireless traffic data in the past p time slots to predict the wireless traffic in the upcoming p time slots, the model is able to capture the relationship between the wireless traffic of the time slots in the last time period and those in the upcoming time period, which is helpful to improve the prediction accuracy. Moreover, by employing temporal aggregation, the prediction errors of the prediction results of the wireless traffic in the upcoming p time slots could offset each other, which could raise the prediction accuracy.

V. MULTI-START SIMULATED ANNEALING ALGORITHM

According to the system model introduced in Section III, in order to minimize the power consumption of the BBU pool, we need to minimize the total number of used BBUs in the BBU pool, as shown in Theorem 1.

Theorem 1: At any time period, minimizing the power consumption of the BBU pool is equivalent to minimizing the total number of used BBUs in the BBU pool.

Proof: According to Equation (1), (2) and (3), we have:

$$\begin{aligned} P_{total}(t) &= \sum_{i=1}^{N} u_i(t) \left(\sum_{k=1}^{6} \lambda_k s_{i,k}(t) + P_{BBU}^{basic}(t) \right) + P_{pool}^{basic}(t) \\ &= \sum_{i=1}^{N} u_i(t) \left(\sum_{k=1}^{6} \lambda_k \sum_{j=1}^{M} C_{i,j}(t) s_{j,k}(t) + P_{BBU}^{basic}(t) \right) \\ &\quad + P_{pool}^{basic}(t) \\ &= \sum_{i=1}^{N} u_i(t) \sum_{k=1}^{6} \lambda_k \sum_{j=1}^{M} C_{i,j}(t) s_{j,k}(t) + \sum_{i=1}^{N} u_i(t) P_{BBU}^{basic}(t) \\ &\quad + P_{pool}^{basic}(t) \\ &= \sum_{k=1}^{6} \lambda_k \sum_{j=1}^{M} \left[\sum_{i=1}^{N} C_{i,j}(t) u_i(t) \right] s_{j,k}(t) + \sum_{i=1}^{N} u_i(t) P_{BBU}^{basic}(t) \\ &\quad + P_{pool}^{basic}(t). \end{aligned}$$

According to constraint C3, $\sum_{i=1}^{N} C_{i,j}(t) u_i(t) = 1, \forall j$. Hence, we have:

$$P_{total}(t) = \sum_{k=1}^{6} \lambda_k \sum_{j=1}^{M} s_{j,k}(t) + \sum_{i=1}^{N} u_i(t) P_{BBU}^{basic}(t) + P_{pool}^{basic}(t).$$

And we can find that $P_{total}(t)$ is consisted of three portions, and the first portion is determined by the wireless traffic of RRHs, which can not be lowered by changing the allocation of computing resource. Moreover, the value of the third portion, $P_{pool}^{basic}(t)$, is a constant. Hence, minimizing $P_{total}(t)$ can only be achieved by minimizing its second portion. In addition, in the second portion, the value of $P_{BBU}^{basic}(t)$ is also a constant. Hence, at time period t, minimizing the power consumption of the BBU pool, $P_{total}(t)$, is equivalent to minimizing $\sum_{i=1}^{N} u_i(t)$, which is the total number of used BBUs in the BBU pool. ■

Hence, the optimization objective could be transformed into the equivalent form:

$$\text{Minimize} \quad \sum_{i=1}^{N} u_i(t),$$

which means that the formulated problem is essentially an offline four-constraint bin packing problem, as the RRHs could be seen as objects and BBUs could be regarded as bins and our aim is to minimize the number of used bins.

Theorem 2: The formulated problem is NP-hard.

The proof of Theorem 2 can be found in Appendix A. Due to the NP-hardness of the formulated problem, there doesn't exist any exact algorithm for finding the optimal solution within an acceptable amount of time. Moreover, the formulated problem is essentially a combinatorial optimization problem, so traditional non-convex optimization methods cannot be applied for solving this problem. For this kind of problem, metaheuristic algorithms are usually utilized for seeking high-quality suboptimal solutions.

In this paper, we propose a Multi-start Simulated Annealing (MSA) algorithm for solving the formulated offline four-constraint bin packing problem. The flowchart of MSA algorithm is shown in Fig. 3.

From Fig. 3, we can find that the design of MSA algorithm is based on Simulated Annealing (SA) algorithm. SA algorithm is a classic metaheuristic algorithm which has good performance on finding high-quality suboptimal solutions by randomly accepting worse solutions in order to overcome getting stuck at local minima. However, the position of the start might have influence on the final result. Hence, in order to better seek for the global minimum, the SA algorithm is run several times with multiple randomly generated starts to mitigate the influence of the position of the start on the final result.

In MSA, First Fit (FF) algorithm is utilized for allocating BBUs to RRHs under the four constraints. In FF algorithm, the objects are put into bins in sequence and one object will be put into the first bin which could accomodate it. Hence, changing the input sequence of RRHs of the FF algorithm

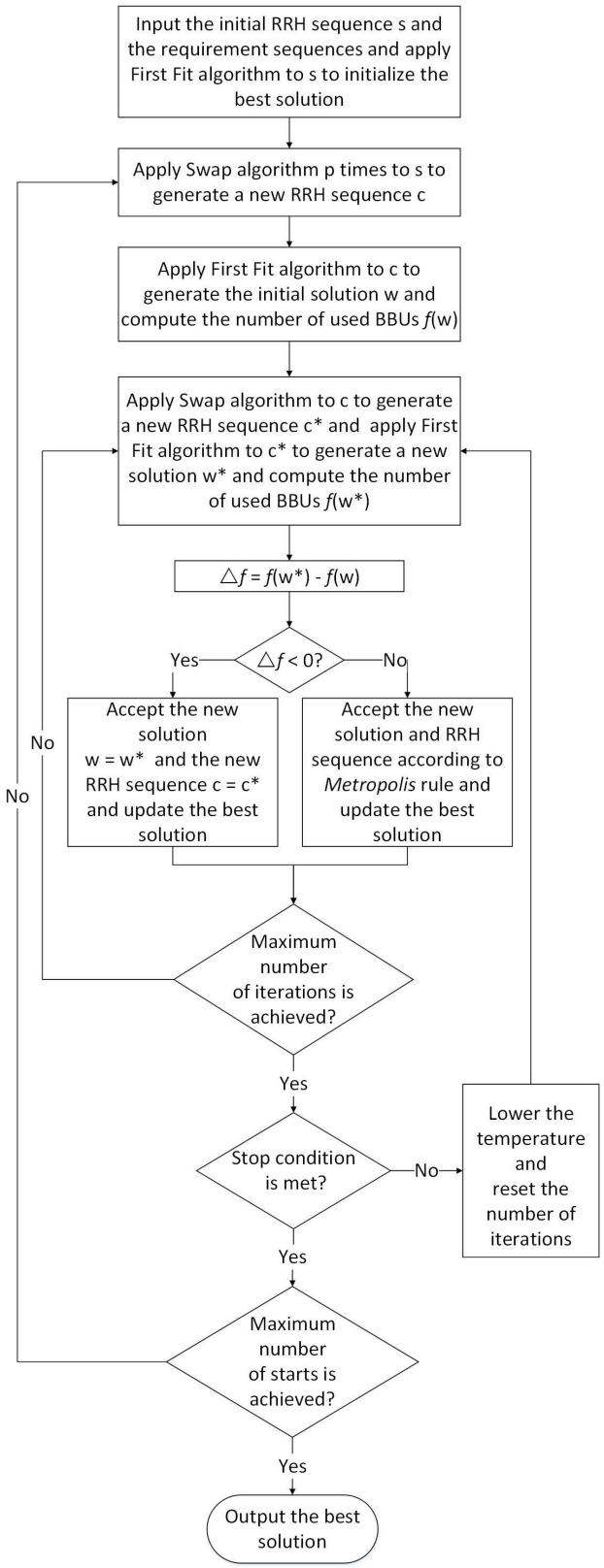


FIGURE 3. The flowchart of MSA algorithm.

might generate new allocation results. In order to generate new starts for the SA algorithm, we apply Swap algorithm a number of times, which exchanges the positions of two RRHs randomly chosen from the input RRH sequence to generate a new RRH sequence. After a new start is generated, it is utilized as the input of SA algorithm. Different from SA algorithm, MSA algorithm keeps a record of the global best solution. Whenever a new solution is accepted for replacing the current solution, it is compared with the global best solution and if it is better, the global best solution will be updated.

In SA algorithm, a new solution is generated by applying Swap algorithm to the current RRH sequence to generate a new RRH sequence, and using the sequence as the input of the FF algorithm. If the performance of the new solution is better than the current solution, which means the new solution uses fewer BBUs, the current solution and RRH sequence will be replaced by the new solution and RRH sequence. If the performance of the new solution is worse than the current solution, the current solution and

Algorithm 1 First Fit Algorithm

Input: The total number of RRHs, M;
The total number of BBUs, N;
The sequence of RRHs, s;
The sequences of uplink and downlink computing resource requirements of RRHs, R_{up}, R_{down};
The sequences of uplink and downlink CPRI bandwidth requirements of RRHs, W_{up}, W_{down};
The maximum uplink and downlink baseband processing capacities of one BBU, R_{up}^{max}, R_{down}^{max};
The maximum uplink and downlink bandwidths of one CPRI, W_{up}^{max}, W_{down}^{max};
Output: The number of used BBUs, n;
The set of RRHs to be served by the j-th BBU, S_j ($j = 1$ to N);
$n = 0$;
for $j = 1$ to N **do**
  $S_j = \emptyset$;
end for
for $i = 1$ to M **do**
  Flag = 0;
  for $j = 1$ to n **do**
    if $\sum_{k \in S_j} R_{up}[k] + R_{up}[s[i]] \leq R_{up}^{max}$
    and $\sum_{k \in S_j} R_{down}[k] + R_{down}[s[i]] \leq R_{down}^{max}$
    and $\sum_{k \in S_j} W_{up}[k] + W_{up}[s[i]] \leq W_{up}^{max}$
    and $\sum_{k \in S_j} W_{down}[k] + W_{down}[s[i]] \leq W_{down}^{max}$ **then**
      $S_j = S_j \cup \{s[i]\}$;
      Flag = 1;
      Break;
    end if
  end for
  if Flag == 0 **then**
    $n = n + 1$;
    $S_n = S_n \cup \{s[i]\}$;
  end if
end for

RRH sequence is replaced by the new solution randomly according to the Metropolis rule, which is as shown below.

$$P(w = w^*) = \exp\left(\frac{f(w) - f(w^*)}{kT}\right), \text{if } f(w^*) > f(w), \quad (10)$$

where T is the current temperature and k is the temperature scale factor. In MSA algorithm, lowering temperature T means that it is less possible for a worse solution to be accepted.

At each temperature, there exists N_I iterations of generating new solutions. After N_I iterations, the algorithm will check whether the stop condition is met, which means the current temperature could not be lowered anymore. If there still exists room for lowering temperature, the temperature will be lowered. At each time, the temperature is lowered by multiplying the current temperature with the cooling rate, r. Hence, by gradually lowering the temperature, according to Equation (10), the probability of accepting worse solutions will be gradually lowered and convergence will eventually be achieved.

Property 1: The time complexity of MSA algorithm is $\mathcal{O}(N_S N_T N_I M \log M)$, where N_S, N_T and N_I correspond to the number of starts, the maximum number of temperature change and the maximum number of iterations at a specific temperature, respectively.

Proof: From Fig. 3 we can find that there exist three loops in MSA algorithm. In each iteration of the first loop, a new start is generated. And there exist N_S iterations in the first loop. In the second loop, the temperature is changed N_T times, which is the maximum number of temperature change, which is:

$$N_T = \left\lceil \log_r\left(\frac{T_{Minimum}}{T_{Initial}}\right)\right\rceil, \quad (11)$$

where $T_{Minimum}$ and $T_{Initial}$ correspond to the minimum temperature and the initial temperature, respectively. In the third loop, the Swap algorithm and First Fit algorithm are applied N_I times to generate new solutions. At any iteration, the time complexity is determined by First Fit algorithm. The time complexity of First Fit algorithm has already been proven to be $\mathcal{O}(M \log M)$ [46]. Hence, the time complexity of MSA algorithm is $\mathcal{O}(N_S N_T N_I M \log M)$. ■

VI. EXPERIMENTAL RESULTS AND ANALYSIS

A. DATASET DESCRIPTION, DATA PREPROCESSING AND OBSERVATION

In this paper, a range of experiments are conducted utilizing the Telecom Italia Big Data Challenge dataset [47]. Specifically, we utilize the Call Detail Records (CDRs) of the city of Milan from November 1, 2013 to January 1, 2014. The area of Milan is divided into 100 × 100 grids, each of size of 235 × 235 square meters. There are 5 types of CDRs in the dataset, which are: incoming call, outgoing call, received SMS, sent SMS and Internet. For the first four types, one CDR is generated each time a user issues or receives a call or SMS. For Internet, one CDR is generated each time an Internet connection starts or ends and the connection lasts for over 15 minutes or the amount of data a user transferred exceeds 5 MB. The temporal interval of the CDRs is 10 minutes.

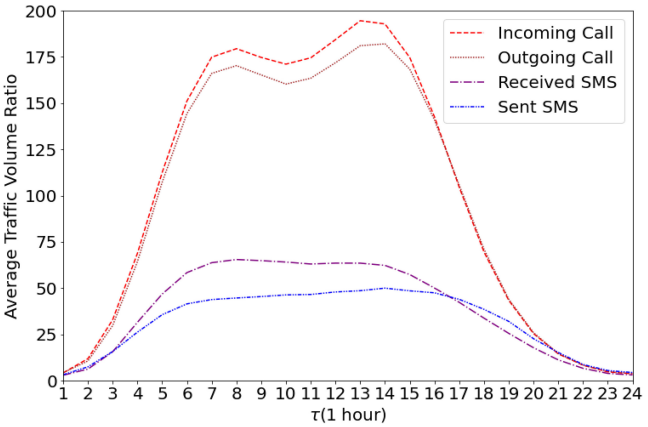


FIGURE 4. ATVR values at 1-hour level.

In this paper, the length of time period utilized for computing resource allocation in CRAN is set to be 1 hour, not 10 minutes. This is because that setting the time period of computing resource allocation to be 10 minutes might make the network unstable and bring excessive overhead. Hence, the original CDRs are aggregated in order to create new CDRs with time interval of 1 hour. However, the original CDRs are also utilized in the training and testing process of our proposed model.

The data shape of the original wireless traffic data of each type of wireless traffic is (8928, 10000). Then in order to utilize temporal aggregation method, the original wireless traffic data is combined and divided into two parts, which are input data and target output data of the model. There exist 8917 pairs of input and target output data, each of size (6, 10000). Hence, the data shape of the input data set and target output data set are both (8917, 6, 10000).

In order to model the temporal correlation in this dataset, the average traffic volume ratio (ATVR) function $\rho(\tau)$ [25] is utilized, which is defined as:

$$\rho(\tau) = \frac{1}{(T-\tau) \times H \times W} \sum_{t=1+\tau}^{T} \sum_{h=1}^{H} \sum_{w=1}^{W} \frac{s_{h,w}(t)}{s_{h,w}(t-\tau)} \quad (12)$$

The ATVR values of four types of wireless traffic at 1-hour level and 10-minute level are plotted in Figs. 4 and 5, respectively.

From Fig. 4, we can find that with the increase of τ from 1 to 8 hours, the ATVR values of the four types of wireless traffic increase dramatically, which indicates that the temporal correlation of wireless traffic in adjacent time periods are stronger than those with longer time gaps within 8 hours. However, we also notice that when the time gap becomes close to 24 hours, the ATVR values decreases. Apparently this is due to the periodicity of wireless traffic data. From Fig. 5, we can find that at 10-minute level, the temporal correlation of wireless traffic in adjacent time periods are

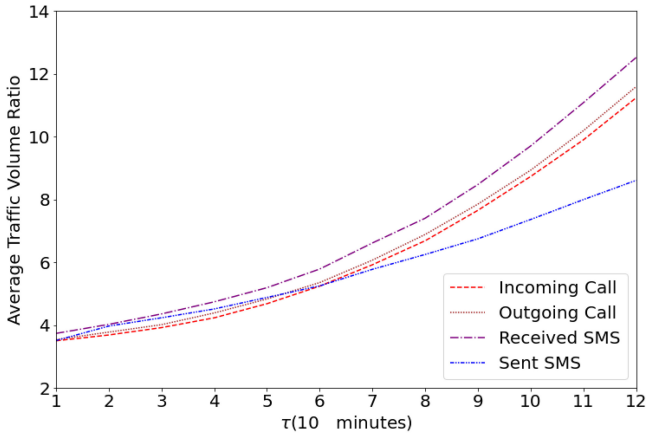


FIGURE 5. ATVR values at 10-minute level.

TABLE 1. Experiment settings of wireless traffic prediction.

Parameter	Value
Size of traffic matrices (H,W)	(100, 100)
Length of one time period	1 hour
Length of one time slot	10 minutes
Number of time slots per time period	6
Ratio of training set to test set	8 : 2

TABLE 2. Parameter settings of two-dimensional CNN LSTM model.

Parameter	Value
Number of filters in the Convolution 2D layer	4
Kernel size of the Convolution 2D layer	(3,3)
Pool size of the Max Pooling 2D layer	(2,2)
Number of units in the LSTM layer	100
Number of units in the Fully Connected layer	10000

stronger than those with longer time gaps within 2 hours. Moreover, by comparing Fig. 4 with 5, we can find that the time correlation at 10-minute level is much stronger than that at 1-hour level, which indicates the possibility of employing the temporal aggregation method to increase the accuracy of wireless traffic prediction.

B. EXPERIMENT SETTINGS AND EVALUATION METRICS OF WIRELESS TRAFFIC PREDICTION

The experiment settings of wireless traffic prediction are given in Table 1. Moreover, the parameter settings of the proposed two-dimensional CNN LSTM model are given in Table 2.

For wireless traffic prediction, three commonly used metrics are adopted for evaluating the prediction performances of the models.

The first metric is Mean Absolute Error (MAE), which is defined as shown below:

$$\mathrm{MAE} = \frac{\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{w=1}^{W}\left|\widehat{s}_{h,w}(t) - s_{h,w}(t)\right|}{T \times H \times W}. \quad (13)$$

The second metric is Root Mean Square Error (RMSE), which is defined as shown below:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(\widehat{s}_{h,w}(t) - s_{h,w}(t)\right)^2}{T \times H \times W}}. \quad (14)$$

The third metric is R-squared (R^2), which is defined as shown below:

$$\mathrm{R}^2 = 1 - \frac{\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(\widehat{s}_{h,w}(t) - s_{h,w}(t)\right)^2}{\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(\overline{s} - s_{h,w}(t)\right)^2}, \quad (15)$$

where $\overline{s}$ is defined as shown below:

$$\overline{s} = \frac{\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{w=1}^{W} s_{h,w}(t)}{T \times H \times W} \quad (16)$$

In order to demonstrate the superiority of our proposed wireless traffic prediction method, a range of experiments are conducted utilizing several state-of-the-art deep learning models for wireless traffic prediction and several variants of LSTM model with the same temporal aggregation approach applied in our proposed model, whose prediction performances are compared with that of our proposed method.

It needs to be pointed out that for our proposed model and the other variants of LSTM model utilized in the experiments, the same data are utilized, whose size of input and target output data are both (8917, 6, 10000) as mentioned before.

However, in order to meet the different requirements on the input data shape of the models, the data are reshaped in different ways. For example, for the proposed two-dimensional CNN LSTM model, the input data shape of the Convolution 2D layer is required to be (100, 100, 1). Hence, for the two-dimensional CNN LSTM model, the input data set shape is reshaped to be (8917, 6, 100, 100, 1). Then according to the ratio of training set to test set, 8:2, the input data set and target output data set are further divided into training and test sets. In the training set, the sizes of input data set and target output data set are (7134, 6, 100, 100, 1) and (7134, 6, 10000), respectively. In the test set, the sizes of input data set and target output data set are (1783, 6, 100, 100, 1) and (1783, 6, 10000), respectively. However, for the Convolution 1D layer of the one-dimensional CNN LSTM model, the input data shape is set to be (2, 10000). Hence, the input data shape is reshaped to be (8917, 3, 2, 10000) for the one-dimensional CNN LSTM model. And in its training set, the sizes of input data set and target output data set are (7134, 3, 2, 10000) and (7134, 6, 10000), respectively. In the test set, the sizes of input data set and target output data set are (1783, 3, 2, 10000) and (1783, 6, 10000), respectively.

C. NUMERICAL RESULTS AND ANALYSIS OF WIRELESS TRAFFIC PREDICTION

The comparison of the prediction performance of our proposed model with those of the baseline models are shown in Tables 3–7, respectively.

TABLE 3. Comparison of prediction performance on incoming call traffic.

Model \ Metric	MAE	RMSE	R^2
LSTM	3.5744	11.5211	0.8570
Bidirectional LSTM	3.3366	12.1612	0.8407
CNN LSTM (1D)	**2.9128**	10.6900	0.8769
CNN LSTM (2D) (Proposed)	3.1683	**9.7747**	**0.8971**
Convolutional LSTM	3.5507	11.7966	0.8501
STCNet [27]	9.6495	18.6455	0.6196
DenseNet [25]	15.4950	27.3147	0.1836
DenseNet -Fusion [25]	9.5138	20.3396	0.5473

TABLE 4. Comparison of prediction performance on outgoing call traffic.

Model \ Metric	MAE	RMSE	R^2
LSTM	4.0499	12.8191	0.8745
Bidirectional LSTM	4.0679	12.9195	0.8725
CNN LSTM (1D)	3.7223	11.9208	0.8915
CNN LSTM (2D) (Proposed)	**3.2975**	**9.2705**	**0.9344**
Convolutional LSTM	4.3968	13.6795	0.8571
STCNet [27]	11.4742	25.5026	0.4957
DenseNet [25]	16.1291	30.1490	0.2952
DenseNet -Fusion [25]	11.9138	24.4430	0.5368

TABLE 5. Comparison of prediction performance on received SMS traffic.

Model \ Metric	MAE	RMSE	R^2
LSTM	6.5730	21.3823	0.7950
Bidirectional LSTM	6.2479	18.8968	0.8399
CNN LSTM (1D)	6.1621	17.8831	0.8566
CNN LSTM (2D) (Proposed)	**6.0979**	**17.0355**	**0.8699**
Convolutional LSTM	6.5923	20.6483	0.8089
STCNet [27]	15.4439	33.3736	0.4967
DenseNet [25]	16.1222	39.3754	0.2994
DenseNet -Fusion [25]	16.1042	36.3581	0.4026

From Tables 3–7, we can find that the proposed model achieves the lowest MAE values on all the five types of traffic except for incoming call and Internet traffic, on which the proposed model achieves the second lowest MAE values. Moreover, the proposed model achieves the lowest RMSE values and the highest R^2 values on all the five different types of wireless traffic. These results demonstrate the superiority of our proposed model over the baseline models on prediction performance.

TABLE 6. Comparison of prediction performance on sent SMS traffic.

Model \ Metric	MAE	RMSE	R^2
LSTM	5.7663	16.4035	0.6420
Bidirectional LSTM	5.2230	12.2920	0.6022
CNN LSTM (1D)	6.4072	17.2712	0.6032
CNN LSTM (2D) (Proposed)	**5.1503**	**15.8464**	**0.6659**
Convolutional LSTM	6.2984	18.7588	0.5319
STCNet [27]	8.8361	20.3771	0.4496
DenseNet [25]	8.8604	21.8691	0.3661
DenseNet -Fusion [25]	8.8245	20.4245	0.4471

TABLE 7. Comparison of prediction performance on Internet traffic.

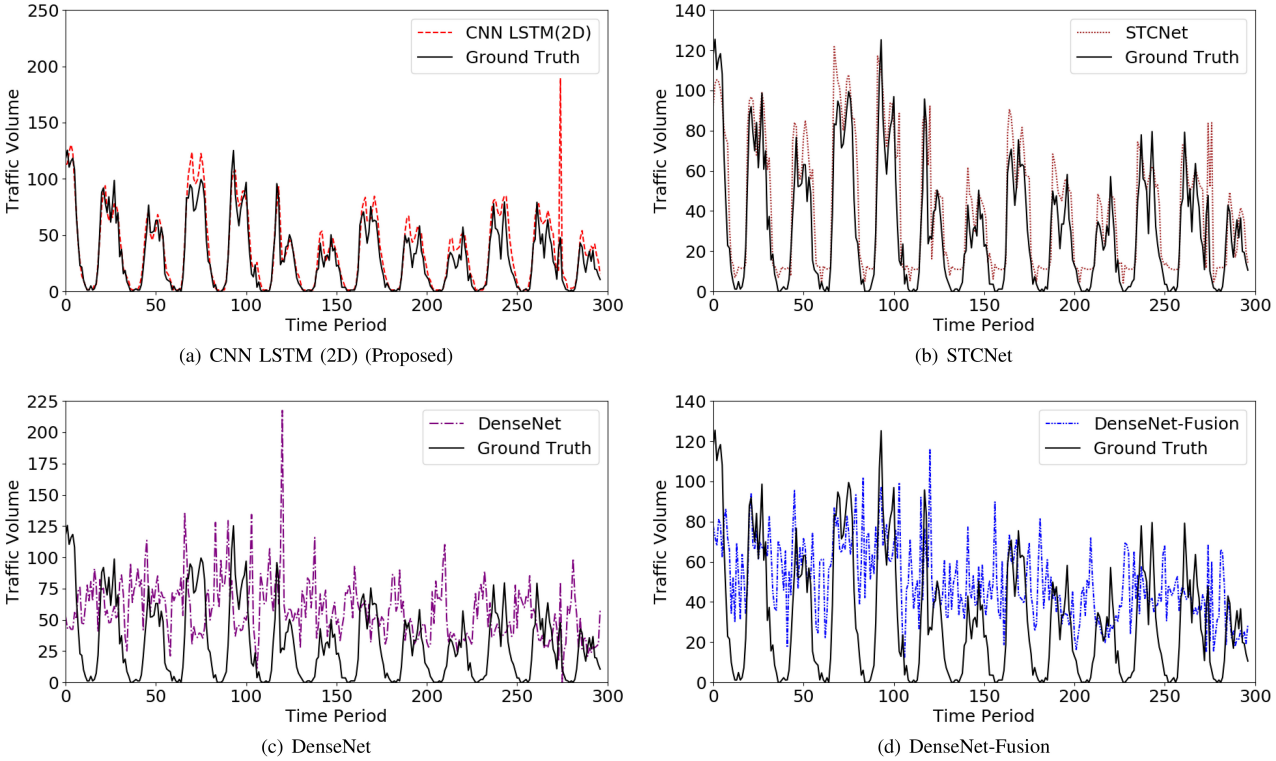
Model \ Metric	MAE	RMSE	R^2
LSTM	69.3479	198.4925	0.8010
Bidirectional LSTM	**55.4445**	154.2303	0.8799
CNN LSTM (1D)	64.9617	160.7744	0.8695
CNN LSTM (2D) (Proposed)	61.7307	**135.5715**	**0.9072**
Convolutional LSTM	93.7289	212.9368	0.7710
STCNet [27]	100.6001	220.2162	0.7521
DenseNet [25]	71.1731	180.2818	0.8338
DenseNet -Fusion [25]	80.2907	220.0689	0.7524

In order to compare the prediction performance of our proposed model with the state-of-the-art models more intuitively, the prediction results of the state-of-the-art deep learning models and target values of the five types of wireless traffic of cell (51, 50) are plotted in Figs. 6–10, respectively. From Figs. 6–10, we can see that generally speaking, the prediction results of our proposed model are closer to the target values than the state-of-the-art models on all of the five types of wireless traffic.

In conclusion, the proposed wireless traffic prediction method based on two-dimensional CNN LSTM model with temporal aggregation could achieve better prediction performance than the baseline models.

D. EXPERIMENT SETTINGS OF ENERGY-EFFICIENT DYNAMIC COMPUTING RESOURCE ALLOCATION IN CRAN

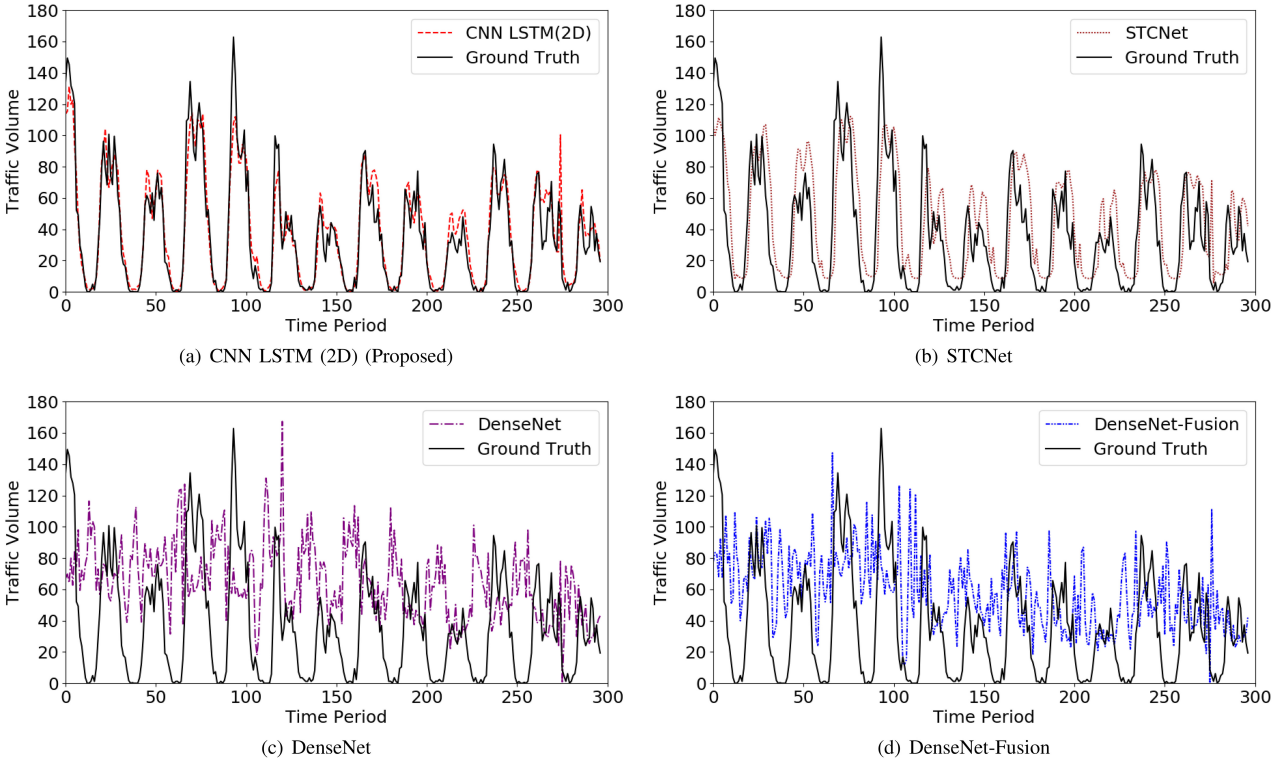
The experiment settings of computing resource allocation in CRAN are given in Table 8.



(a) CNN LSTM (2D) (Proposed) (b) STCNet

(c) DenseNet (d) DenseNet-Fusion

FIGURE 6. Comparison between prediction results and ground truth values of incoming call traffic of cell (51, 50).

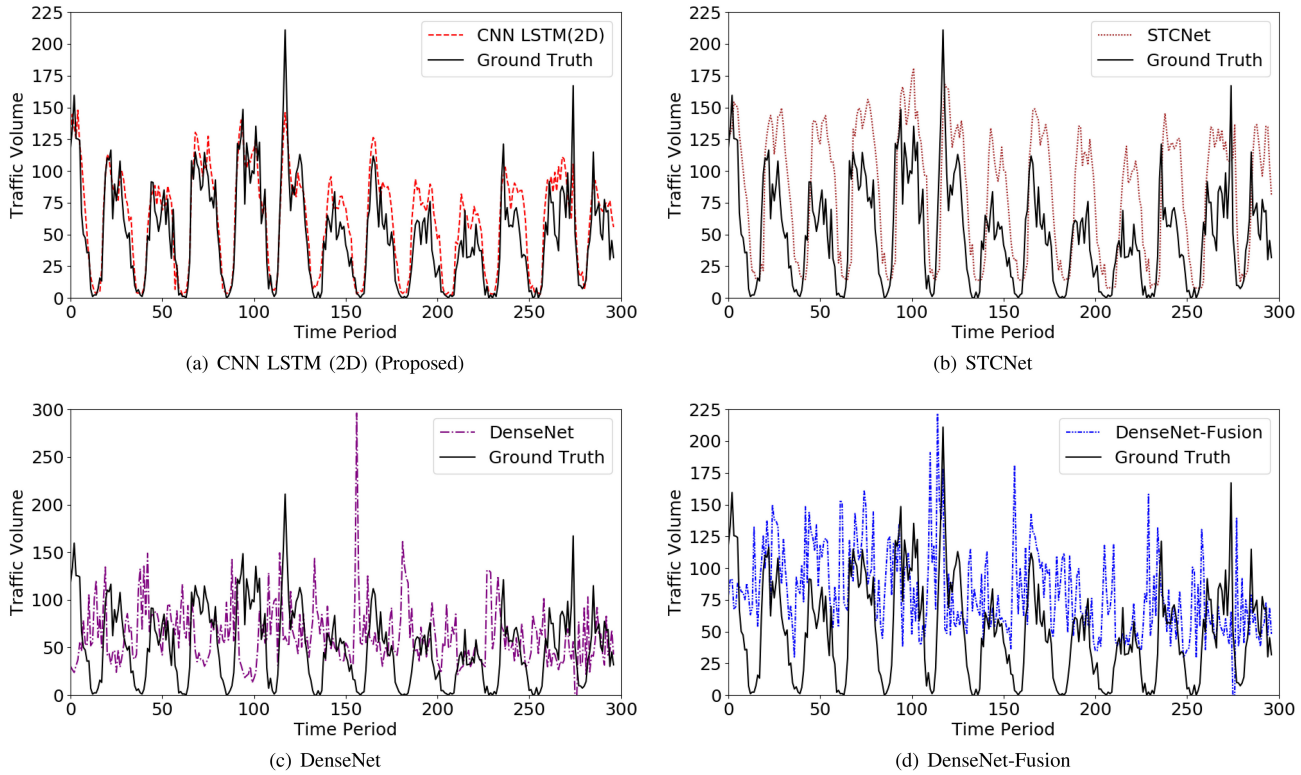


(a) CNN LSTM (2D) (Proposed) (b) STCNet

(c) DenseNet (d) DenseNet-Fusion

FIGURE 7. Comparison between prediction results and ground truth values of outgoing call traffic of cell (51, 50).

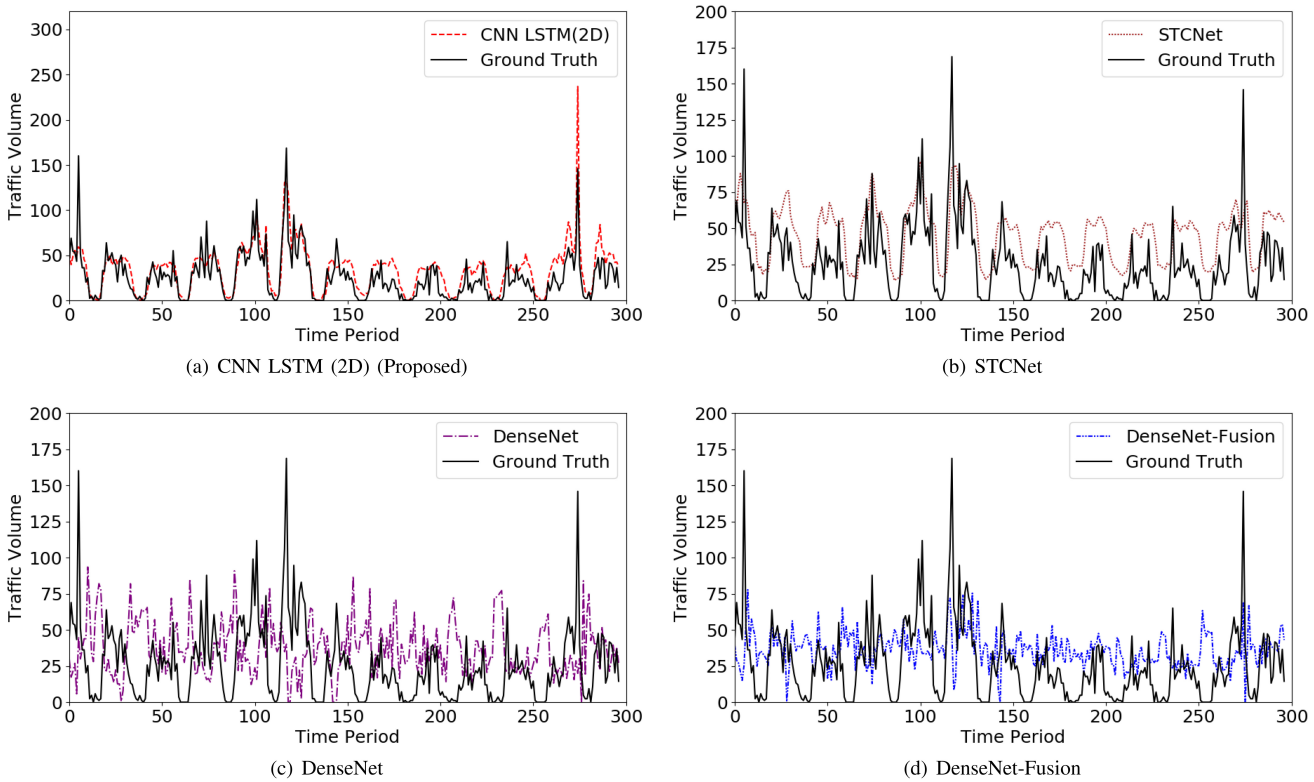
For computing resource allocation, the wireless traffic prediction results of the proposed method are utilized. Due to the large size of the dataset, we only use the wireless traffic prediction results of the central 100 cells for conducting experiments. Moreover, because the wirless traffic of uplink and downlink Internet transmission are not distinguished in the original dataset, we assume that the ratio of uplink Internet wireless traffic



(a) CNN LSTM (2D) (Proposed) (b) STCNet

(c) DenseNet (d) DenseNet-Fusion

FIGURE 8. Comparison between prediction results and ground truth values of received SMS traffic of cell (51, 50).



(a) CNN LSTM (2D) (Proposed) (b) STCNet

(c) DenseNet (d) DenseNet-Fusion

FIGURE 9. Comparison between prediction results and ground truth values of sent SMS traffic of cell (51, 50).

to downlink Internet wireless traffic is 1:2. The parameter settings of our proposed MSA algorithm is given in Table 9.

In order to demonstrate the superiority of our proposed MSA algorithm, several baseline algorithms are implemented and their performances are compared with that of MSA

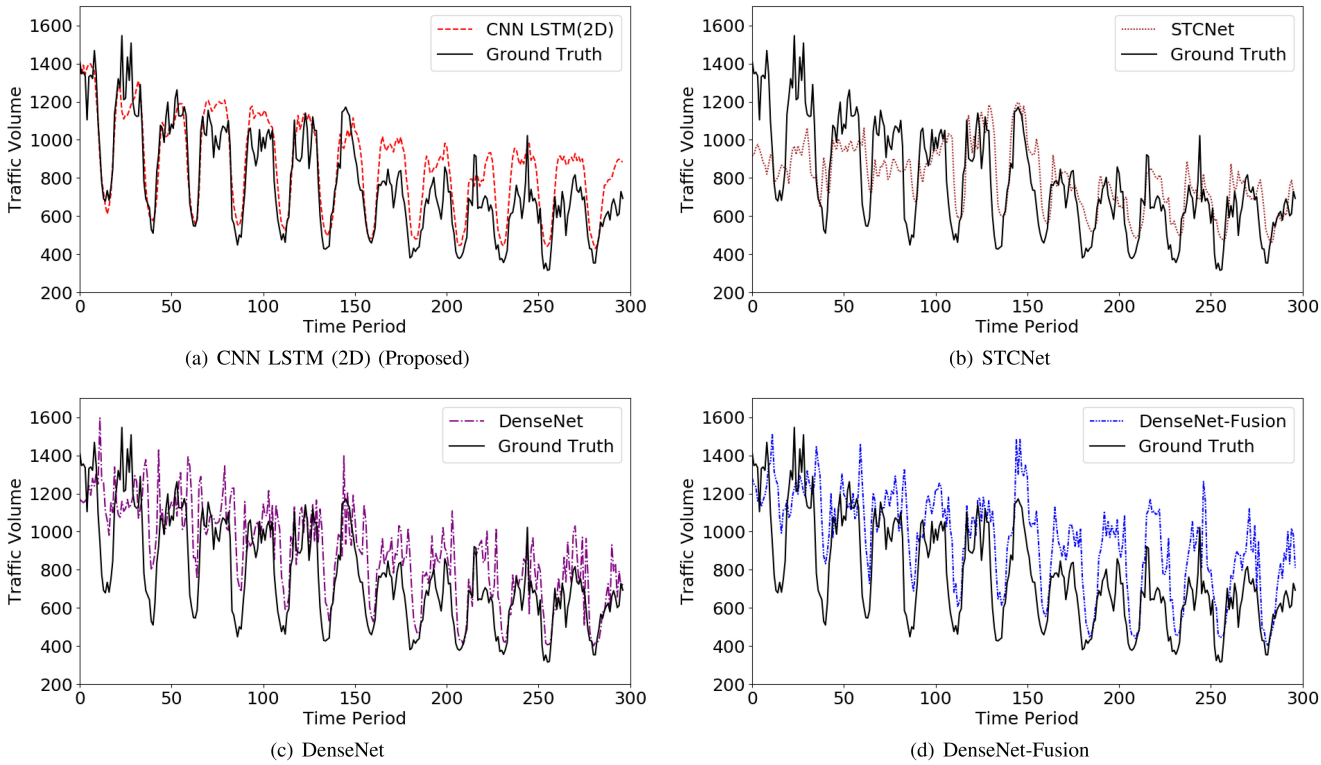


FIGURE 10. Comparison between prediction results and ground truth values of Internet traffic of cell (51, 50).

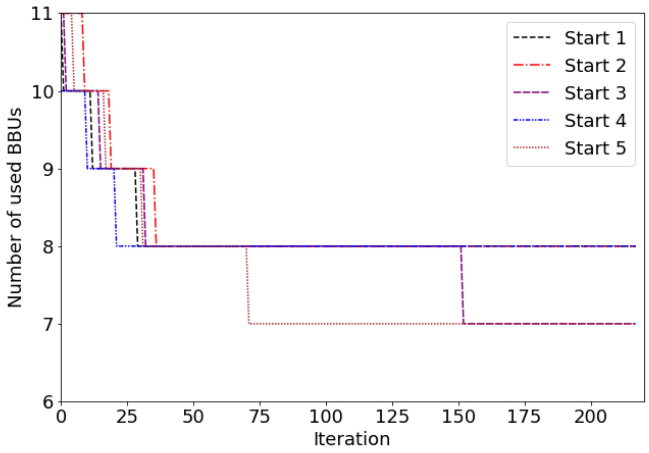


FIGURE 11. Simulation results of MSA algorithm in time period 16.

algorithm. The baseline algorithms are First Fit (FF) algorithm, Genetic Algorithm (GA), Particle Swarm Optimization (PSO) algorithm and Simulated Annealing (SA) algorithm.

E. NUMERICAL RESULTS AND ANALYSIS OF ENERGY-EFFICIENT DYNAMIC COMPUTING RESOURCE ALLOCATION IN CRAN

In order to demonstrate the convergence of MSA algorithm, the simulation results of MSA algorithm in time period 16 are taken as an example, which is plotted in Fig. 11. In Fig. 11, during each iteration, the temperature stays unchanged. And the temperature gradually decreases with the increase of the number of iterations. From 11, we can find that the number of used BBUs gradually converge with the increase of the number of iterations. This is because when the temperature gradually decreases, the probability of accepting a worse solution decreases according to the Metropolis rule. Moreover, we can find that for starts 1, 2 and 4, convergence of the number of used BBUs is achieved at 8. However, for starts 3 and 5, convergence of the number of used BBUs is achieved at 7. This demonstrates that the position of the start might influence the final result. Hence, randomly choosing multiple different starts might generate better results compared with using one single start.

The number of used BBUs of MSA algorithm and the baseline algorithms are plotted in Fig. 12. From Fig. 12, we can find that almost in every time period, the MSA algorithm generates the minimum number of used BBUs.

The total power consumption of BBU pool of the MSA algorithm and the baseline algorithms are compared in Fig. 13. In Fig. 13, the fixed algorithm means that the BBU allocation is not changed during all the periods and we assume that the maximum number of BBUs are used. From Fig. 13, we can find that the fixed algorithm consumes the highest total power consumption of the BBU pool, which is about twice as much as the total power consumption of the BBU pool of any other dynamic computing resource allocation algorithm. This demonstrates the usefulness of

TABLE 8. Experiment settings of energy-efficient dynamic computing resource allocation in CRAN.

Parameter	Value
Number of RRHs	100
Number of BBUs	25
Number of periods	297
Length of one time period	1 hour
Basic power consumption of the BBU pool	30 W
Basic power consumption of one BBU	300 W
Maximum power consumption of one BBU	330 W
Maximum uplink baseband signal processing capacity of one BBU	200 GOPS
Maximum downlink baseband signal processing capacity of one BBU	200 GOPS
Maximum uplink bandwidth of one CPRI	10 Gbps
Maximum downlink bandwidth of one CPRI	10 Gbps
Coefficient of computing resource requirement of uplink/downlink call traffic	0.016
Coefficient of computing resource requirement of uplink/downlink SMS traffic	4×10^{-7}
Coefficient of computing resource requirement of uplink/downlink Internet traffic	0.02
Coefficient of CPRI bandwidth requirement of uplink/downlink call traffic	0.0006
Coefficient of CPRI bandwidth requirement of uplink/downlink SMS traffic	3×10^{-8}
Coefficient of CPRI bandwidth requirement of uplink/downlink Internet traffic	0.001
Coefficient of power consumption of uplink/downlink call traffic	0.0012
Coefficient of power consumption of uplink/downlink SMS traffic	3×10^{-8}
Coefficient of power consumption of uplink/downlink Internet traffic	0.0015

TABLE 9. Parameter settings of MSA algorithm.

Parameter	Value
Number of starts	5
Number of times of applying Swap algorithm to generate a new start	100
Number of iterations at each temperature	10000
Initial temperature	50
Temperature scale factor	0.01
Cooling rate	0.97
Minimum temperature	0.1

dynamic computing resource allocation for reducing the total power consumption of the BBU pool. Moreover, we can find that the proposed MSA algorithm achieves the lowest total power consumption of the BBU pool, which demonstrate the superiority of MSA algorithm over the baseline

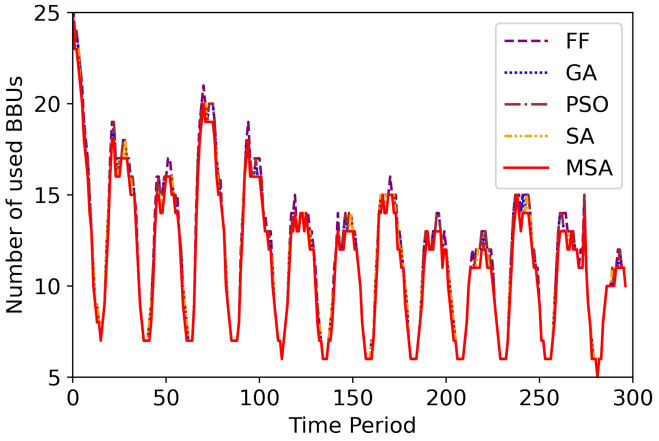


FIGURE 12. Comparison of number of used BBUs.

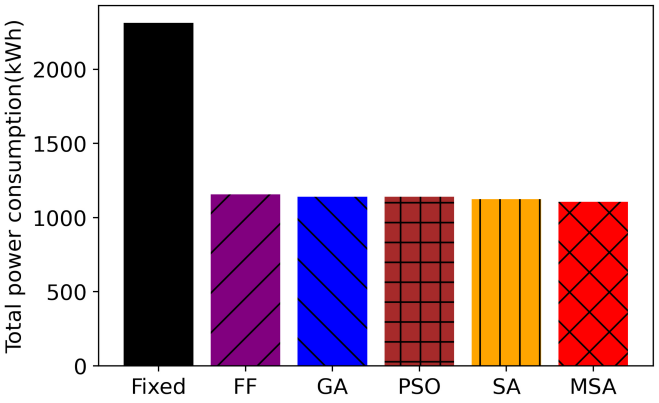


FIGURE 13. Comparison of total power consumption of BBU pool.

algorithms on reducing the total power consumption of the BBU pool.

VII. CONCLUSION

In this paper, the problem of energy-efficient dynamic computing resource allocation in CRAN is investigated. Aiming at reducing the power consumption of the BBU pool, the problem of energy-efficient dynamic computing resource allocation in CRAN is formulated as an offline four-constraint bin packing problem, taking both the uplink and downlink baseband signal processing capacities of BBUs and both the uplink and downlink CPRI bandwidths into consideration. The demands on uplink and downlink computing resources and CPRI bandwidths of RRHs are estimated according to the prediction results of wireless traffic of RRHs. For wireless traffic prediction, a novel method based on two-dimensional CNN LSTM model with temporal aggregation is proposed. For solving the formulated offline four-constraint bin packing problem, a Multi-start Simulated Annealing (MSA) algorithm is proposed. Extensive experiments are conducted utilizing a real-world dataset, whose results demonstrate that the proposed wireless traffic prediction method could outperform the state-of-the-art deep learning models on prediction performance and the

proposed MSA algorithm could achieve lower power consumption of the BBU pool compared with the state-of-the-art heuristic algorithms.

APPENDIX A
PROOF OF THEOREM 2

We prove this theorem through reducing the offline one-dimensional bin packing problem [48], which has already been proven to be NP-hard, to the formulated problem.

The offline one-dimensional bin packing problem is defined as follows.

Definition 1 (Offline One-Dimensional Bin Packing Problem): Given a set of items, which is denoted as $\mathcal{N} = \{1, \ldots, N\}$, with the weight set, which is denoted as $\mathcal{W} = \{w_1, \ldots, w_N\}$, where w_i denotes the weight of the item i. The aim is to pack the items into bins with capacity C and minimize the number of used bins without violating the capacity constraints.

We can reduce the offline one-dimensional bin packing problem to the formulated problem as follows. The item set $\mathcal{N}$ can be regarded as the RRH set in the formulated problem. Moreover, the weight of the item i, w_i, can be seen as the uplink computing resource requirement of the i-th RRH. Besides, we regard the capacity of bins, C, as the maximum uplink baseband processing capacity of a single BBU, R_{up}^{max}. In addition, we set the downlink computing resource requirement, the uplink and downlink CPRI bandwidth requirements of RRHs to be 0s. And we set the maximum downlink baseband processing capacity of one BBU, the maximum uplink and downlink bandwidths of one CPRI to be 0s.

Then the optimal solution of the formulated problem could be transformed into the optimal solution of the offline one-dimensional bin packing problem, by regarding the i-th RRH as the i-th item and BBUs as bins. Hence, the offline one-dimensional bin packing problem can be solved by solving the formulated problem. Therefore, the offline one-dimensional bin packing problem is reduced to the formulated problem in polynomial time. This means that the formulated problem is also NP-hard.

REFERENCES

[1] T. Hatt and E. Kolta, "5G Energy Efficiencies: Green Is the New Black." GSMA Intelligence. Nov. 2020. [Online]. Available: https://data.gsmaintelligence.com/research/research/research-2020/5g-energy-efficiencies-green-is-the-new-black

[2] M. Walker. "Operators Facing Power Cost Crunch." MTN Consulting. Mar. 2020. [Online]. Available: https://www.mtnconsulting.biz/product/operators-facing-power-cost-crunch/

[3] Z. Wang, D. W. K. Ng, V. W. S. Wong, and R. Schober, "Transmit beamforming for QoE improvement in C-RAN with mobile virtual network operators," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, 2016, pp. 1–6.

[4] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016.

[5] S. S. Pappas, L. Ekonomou, D. C. Karamousantas, G. E. Chatzarakis, S. K. Katsikas, and P. Liatsis, "Electricity demand loads modeling using AutoRegressive Moving Average (ARMA) models," *Energy*, vol. 33, no. 9, pp. 1353–1360, Sep. 2008.

[6] J. Contreras, R. Espinola, F. J. Nogales, and A. J. Conejo, "ARIMA models to predict next-day electricity prices," *IEEE Trans. Power Syst.*, vol. 18, no. 3, pp. 1014–1020, Aug. 2003.

[7] E. Z. Martinez, E. A. S. D. Silva, and A. L. D. Fabbro, "A SARIMA forecasting model to predict the number of cases of dengue in Campinas, State of São Paulo, Brazil," *Revista da Sociedade Brasileira de Medicina Tropical*, vol. 44, no. 4, pp. 436–440, Aug. 2011.

[8] T. Van Gestel *et al.*, "Financial time series prediction using least squares support vector machines within the evidence framework," *IEEE Trans. Neural Netw.*, vol. 12, no. 4, pp. 809–821, Jul. 2001.

[9] F. Martínez, M. P. Frías, M. D. Pérez, and A. J. Rivera, "A methodology for applying *k*-nearest neighbor to time series forecasting," *Artif. Intell. Rev.*, vol. 52, no. 3, pp. 2019–2037, Nov. 2019.

[10] H. I. Erdal and O. Karakurt, "Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms," *J. Hydrol.*, vol. 477, pp. 119–128, Jan. 2013.

[11] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," Aug. 2017. *arXiv:1704.02971*.

[12] W. Jiang and H. D. Schotten, "Deep learning for fading channel prediction," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 320–332, 2020.

[13] Y. Tian, G. Pan, and M.-S. Alouini, "Applying deep-learning-based computer vision to wireless communications: Methodologies, opportunities, and challenges," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 132–143, 2020.

[14] Y. Wang, W. Liao, and Y. Chang, "Gated recurrent unit network-based short-term photovoltaic forecasting," *Energies*, vol. 11, no. 8, p. 2163, Aug. 2018.

[15] M. Xu, Y. Dong, Z. Li, M. Han, and T. Xing, "A novel time series prediction model based on deep sparse autoencoder," in *Proc. 37th IEEE Chin. Control Conf.*, Wuhan, China, 2018, pp. 1678–1682.

[16] R. Hrasko, A. G. C. Pacheco, and R. A. Krohling, "Time series prediction using restricted Boltzmann machines and backpropagation," *Procedia Comput. Sci.*, vol. 55, pp. 990–999, Jul. 2015.

[17] Y. Cheng, X. Zhou, S. Wan, and K.-K. R. Choo, "Deep belief network for meteorological time series prediction in the Internet of Things," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4369–4376, Jun. 2018.

[18] A. Silvestrini and D. Veredas, "Temporal aggregation of univariate and multivariate time series models: A survey," *J. Econ. Surv.*, vol. 22, no. 3, pp. 458–497, Jul. 2008.

[19] E. C. Man Jr., M. Garey, and D. Johnson, "Approximation algorithms for bin packing: A survey," in *Approximation Algorithms for NP-hard Problems*. Boston, MA, USA: PWS Publ., Aug. 1996, pp. 46–93.

[20] E. G. Coffman, J. Csirik, G. Galambos, S. Martello, and D. Vigo, "Bin packing approximation algorithms: Survey and classification," in *Handbook of Combinatorial Optimization*. New York, NY, USA: Springer, Jan. 2013, pp. 455–531.

[21] Y. Jiang, Z. Cao, and J. Zhang, "Solving 3D bin packing problem via multimodal deep reinforcement learning," in *Proc. 20th Int. Conf. Auton. Agents MultiAgent Syst.*, London, U.K., 2021, pp. 1548–1550.

[22] R. L. Rao and S. S. Iyengar, "Bin-packing by simulated annealing," *Comput. Math. Appl.*, vol. 27, no. 5, pp. 71–82, Mar. 1994.

[23] H. Iima and T. Yakawa, "A new design of genetic algorithm for bin packing," in *Proc. IEEE Congr. Evol. Comput.*, vol. 2. Canberra, ACT, Australia, 2003, pp. 1044–1049.

[24] D. Liu, K. C. Tan, C. K. Goh, and W. K. Ho, "On solving multiobjective bin packing problems using particle swarm optimization," in *Proc. IEEE Int. Conf. Evol. Comput.*, Vancouver, BC, Canada, 2006, pp. 2095–2102.

[25] 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.

[26] C. Zhang and P. Patras, "Long-term mobile traffic forecasting using deep spatio-temporal neural networks," in *Proc. 18th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2018, pp. 231–240.

[27] 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.

[28] Q. He, A. Moayyedi, G. Dán, G. Koudouridis, and P. Tengkvist, "A meta-learning scheme for adaptive short-term network traffic prediction," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 10, pp. 2271–2283, Oct. 2020.

[29] Q. Zeng, Q. Sun, G. Chen, H. Duan, C. Li, and G. Song, "Traffic prediction of wireless cellular networks based on deep transfer learning and cross-domain data," *IEEE Access*, vol. 8, pp. 172387–172397, 2020.
[30] N. Zhao, Z. Ye, Y. Pei, Y.-C. Liang, and D. Niyato, "Spatial-temporal attention-convolution network for citywide cellular traffic prediction," *IEEE Commun. Lett.*, vol. 24, no. 11, pp. 2532–2536, Nov. 2020.
[31] S. Zhao *et al.*, "Cellular network traffic prediction incorporating handover: A graph convolutional approach," in *Proc. 17th Annu. IEEE Int. Conf. Sens. Commun. Netw. (SECON)*, Como, Italy, 2020, pp. 1–9.
[32] C. Zhang, S. Dang, B. Shihada, and M.-S. Alouini, "Dual attention-based federated learning for wireless traffic prediction," in *Proc. IEEE Conf. Comput. Commun.*, Vancouver, BC, Canada, 2021, pp. 1–10.
[33] S. P. Sone, J. J. Lehtomäki, and Z. Khan, "Wireless traffic usage forecasting using real enterprise network data: Analysis and methods," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 777–797, 2020.
[34] W. Wang, C. Zhou, H. He, W. Wu, W. Zhuang, and X. S. Shen, "Cellular traffic load prediction with LSTM and Gaussian process regression," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Dublin, Ireland, 2020, pp. 1–6.
[35] X. Xing, Y. Lin, H. Gao, and Y. Lu, "Wireless traffic prediction with series fluctuation pattern clustering," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Montreal, QC, Canada, 2021, pp. 1–6.
[36] B. J. Sahu, S. Dash, N. Saxena, and A. Roy, "Energy-efficient BBU allocation for green C-RAN," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1637–1640, Jul. 2017.
[37] E. Aqeeli, A. Moubayed, and A. Shami, "Power-aware optimized RRH to BBU allocation in C-RAN," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1311–1322, Feb. 2018.
[38] M. M. Abdelhakam, M. M. Elmesalawy, M. K. Elhattab, and H. H. Esmat, "Energy-efficient BBU pool virtualisation for C-RAN with quality of service guarantees," *IET Commun.*, vol. 14, no. 1, pp. 11–20, Nov. 2019.
[39] M. Barahman, L. M. Correia, and L. S. Ferreira, "A QoS-demand-aware computing resource management scheme in cloud-RAN," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 1850–1863, 2020.
[40] W.-C. Chien, C.-F. Lai, and H.-C. Chao, "Dynamic resource prediction and allocation in C-RAN with edge artificial intelligence," *IEEE Trans. Ind. Informat.*, vol. 15, no. 7, pp. 4306–4314, Jul. 2019.
[41] M. Mouawad, Z. Dziong, and M. Khan, "Quality of Service aware dynamic BBU-RRH mapping based on load prediction using Markov model in C-RAN," in *Proc. IEEE Int. Conf. Internet Things (iThings) IEEE Green Comput. Commun. (GreenCom) IEEE Cyber Phys. Soc. Comput. (CPSCom) IEEE Smart Data (SmartData)*, Halifax, NS, Canada, 2018, pp. 1907–1912.
[42] M. Khan, R. S. Alhumaima, and H. S. Al-Raweshidy, "Quality of Service aware dynamic BBU-RRH mapping in cloud radio access network," in *Proc. Int. Conf. Emerg. Technol. (ICET)*, Peshawar, Pakistan, 2015, pp. 1–5.
[43] F. Zhang, J. Zheng, Y. Zhang, and L. Chu, "An efficient and balanced BBU computing resource allocation algorithm for cloud radio access networks," in *Proc. IEEE 85th Veh. Technol. Conf. (VTC Spring)*, Sydney, NSW, Australia, 2017, pp. 1–5.
[44] M. Khan, F. A. Sabir, and H. S. Al-Raweshidy, "Load balancing by dynamic BBU-RRH mapping in a self-optimised cloud radio access network," in *Proc. IEEE 24th Int. Conf. Telecommun. (ICT)*, Limassol, Cyprus, 2017, pp. 1–5.
[45] M. Y. Lyazidi, N. Aitsaadi, and R. Langar, "Dynamic resource allocation for cloud-RAN in LTE with real-time BBU/RRH assignment," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, 2016, pp. 1–6.
[46] E. G. Coffman Jr., M. R. Garey, and D. S. Johnson, "Approximation algorithms for bin-packing—An updated survey," in *Algorithm Design For Computer System Design*. Vienna, Austria: Springer, 1984, pp. 49–106.
[47] G. Barlacchi *et al.*, "A multi-source dataset of urban life in the city of Milan and the Province of Trentino," *Sci. Data*, vol. 2, no. 1, pp. 1–15, Oct. 2015.
[48] A. C. F. Alvim, C. C. Ribeiro, F. Glover, and D. J. Aloise, "A hybrid improvement heuristic for the one-dimensional bin packing problem," *J. Heuristics*, vol. 10, no. 2, pp. 205–229, Mar. 2004.

YONGQIN FU (Graduate Student Member, IEEE) received the B.E. degree in information engineering from Southeast University, Nanjing, China, in 2015, and the M.E. degree in computer science and technology from Zhejiang University, Hangzhou, China, in 2019. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Western University, London, ON, Canada.

His current research interests include customized and application-oriented resource allocation in 5G beyond and 6G networks, and sensor data analysis in Internet of Things systems.

XIANBIN WANG (Fellow, IEEE) received the Ph.D. degree in electrical and computer engineering from the National University of Singapore in 2001.

He is a Professor and a Tier-1 Canada Research Chair with Western University, Canada. Prior to joining Western, he was with Communications Research Centre Canada (CRC) as a Research Scientist/Senior Research Scientist from July 2002 to December 2007. From January 2001 to July 2002, he was a System Designer with STMicroelectronics. His current research interests include 5G/6G technologies, Internet of Things, communications security, machine learning, and intelligent communications.

Prof. Wang has received many awards and recognitions, including the Canada Research Chair, the CRC President's Excellence Award, the Canadian Federal Government Public Service Award, the Ontario Early Researcher Award, and six IEEE Best Paper Awards. He has over 500 highly cited journal and conference papers, in addition to 30 granted and pending patents and several standard contributions. He currently serves/has served as the editor-in-Chief, an associate editor-in-chief, an editor/associate editor for over ten journals. He was involved in many IEEE conferences, including GLOBECOM, ICC, VTC, PIMRC, WCNC, CCECE, and CWIT, in different roles, such as a General Chair, a Symposium Chair, a Tutorial Instructor, a Track Chair, a Session Chair, a TPC Co-Chair, and a Keynote speaker. He has been nominated as an IEEE Distinguished Lecturer several times during the last ten years. He is currently serving as the Chair of IEEE London Section and ComSoc Signal Processing and Computing for Communications Technical Committee. He is a Fellow of Canadian Academy of Engineering and Engineering Institute of Canada, and an IEEE Distinguished Lecturer.