LSTM based Proactive Access Point Selection and Mobility Load Balancing for Ultra-Dense Networks

Munazza Shabbir, Sithamparanathan Kandeepan, Akram Al-Hourani and Wayne Rowe.

***Abstract*—To cater for the growing demand of capacity, Access Point (AP) densification is a promising solution. Within an Ultra Dense Network (UDN), a mobile User Equipment (UE) can associate with one of many candidate access points. Association with an optimal AP, from multiple eligible APs in proximity of a UE, while avoiding unbalanced UE distribution, becomes a challenge. Legacy Mobility Load Balancing (MLB) methods are usually reactive by design which hinders their efficiency to solve time restrained MLB problem. In this paper, an improved MLB framework is proposed, which utilises Recurrent Neural Network with Long Short Term Memory (RNN-LSTM), for users' temporal and spatial mobility prediction. It predicts AP's future load based on predicted distribution of UEs and finally optimises network's load in advance via a load aware AP to UE association based Load Balancing (LB) technique. This MLB framework optimises load pre-emptively by visualising congestion beforehand. The simulation results suggest that the performance of our MLB framework is superior to other existing algorithms that address time-constrained LB problem incurred by high mobility.**

***Index Terms*—5G and beyond, ultra-dense networks, mobility load balancing, mobility prediction, user association, Recurrent Neural Networks (RNNs), small cells, Long Short-Term Memory (LSTM) Network.**

I. Introduction

To achieve goals of higher data rates and better coverage, dense deployment of APs has been emerging as a promising technology for 5G/B5G networks. UDN comprises high density of Small Cells (SCs), where a user may receive signals from multiple suitable for association candidate APs simultaneously, due to their proximity. Similarly, mobile UEs will cross cells even more frequently, associating with different APs. Under this dense and complex infrastructure, load balancing among neighbouring APs, while maintaining a good signal quality for UEs, becomes a challenge [1]. Load imbalance is more frequent in SC network due to UE mobility and constrained coverage/resources of SCs, which reduces the network throughput and handover success rate.

In this pretext, selection of ideal AP to associate with, such that Quality of Experience (QoE) requirements of mobile UEs are satisfied and resources of APs are efficiently utilized is critical. Traditional, maximum received signal strength based AP to UE association mechanism, yields unbalanced distribution of users among candidate APs with similar resources, as it does not take load of the APs in consideration in association decisions. Also, conventional MLB techniques are reactive in nature and usually utilise current network statistics for decision making. However, in dynamically changing cell environment where UEs are highly mobile, by the time the optimal network parameters are configured to curb overloading, they become stale. Especially in 5G/B5G networks, where low latency of the algorithm is more demanding, it calls for re-evaluation of AP selection and LB techniques. Therefore, this research article concentrates on proposing a solution focused on proactive load aware AP selection scheme assisted by mobility prediction, for high mobility scenarios.

The authors are with the School of Engineering, RMIT University, Melbourne, Australia. Emails:{munazza.shabbir, kandeepan.sithamparanathan, wayne.rowe, akram.hourani}@rmit.edu.au.

A. Related work

In dense AP deployment, the chance of associating with an unideal target AP increases, when conventional handover/association strategies, relying on the strongest signal received at the UE from surrounding APs are used [2].

Mobility prediction has also been utilised for MLB in some recent research work. Authors in [3] used a Multi-graph Convolutional Network (MGCN) and Gated Recurrent Unit network (GRU) for user's location prediction. Then LB is performed by manipulating Cell Individual Offset (CIO) value for each cell. In [4] authors proposed a MLB framework "OPERA", in which Semi-Markov model was used for mobility prediction, while concurrently optimising coverage and capacity of network. However, one of the limitations is that, in [4] only human walk-based mobility model is implemented and high mobility scenario such as vehicular mobility is not considered. Also, semi-Markov models have limited memory, therefore long-term sequence dependencies in the trajectory data are not recorded. This adversely affects the prediction accuracy and hence performance of LB algorithm. In [5] authors first proposed a forward-looking LB scheme, by incorporating mobility and future cell load prediction, via Bayesian Additive Regression Trees (BART) model. Based on predicted load of cells, authors then propose a LB algorithm which adjusts CIO to avoid overloading.

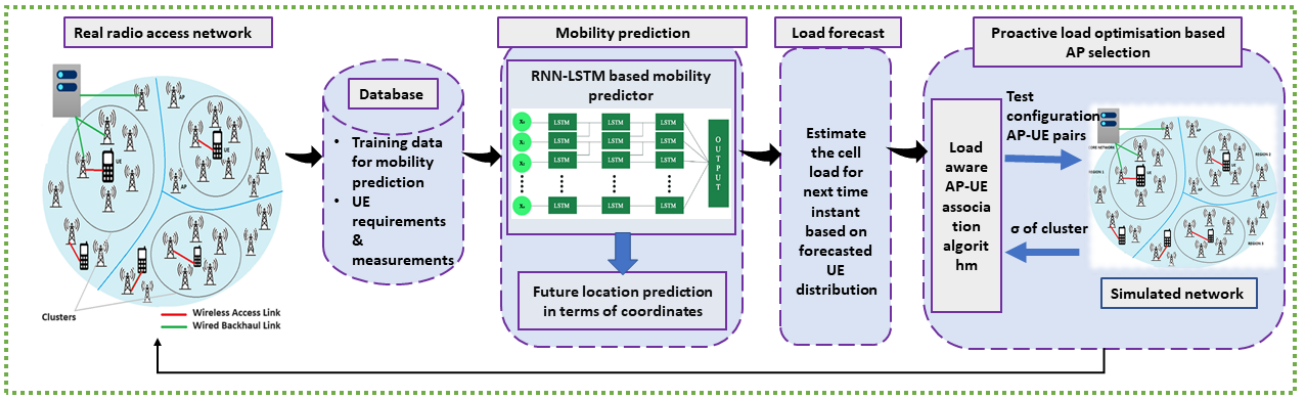

Fig. 1. Proposed proactive MLB framework, comprised of, 5G RAN, mobility prediction, load forecast and load optimisation modules. .

In [6], authors presented Load Balanced Handover Minimized User Association (HMUA-LB) problem. HMUA-LB aimed to minimise number of overloaded APs in the network and to decrease number of handovers experienced by a UE. To solve this problem, a greedy heuristic algorithm LB-USSL is proposed. However, for performance evaluation of LB-USSL, UEs are mobilised at a constant speed of 60 km/hr during simulation, which limits the applicability of the proposed scheme.

B. Contribution

In perspective of aforementioned limitations, we propose a proactive MLB framework powered by RNN-LSTM based mobility prediction (Fig. 1). The contributions of this paper are as follows:

1) In the proposed proactive MLB framework, mobility prediction of UEs with respect to time and location, based on RNN-LSTM is utilised to anticipate loads of SCs. Based on historical trajectory data of UEs, their future locations in terms of coordinates are predicted. Considering anticipated location of all the UEs, the APs most likely to be chosen by these UEs to associate with, can also be predicted. This enables us to visualise future load of all APs as well.

2) Based on predicted utilization of AP resources, load optimization is performed. The proposed load optimization algorithm is an extension of our previous work in [7] and is based on *load aware AP-UE association*. The proactivity of proposed MLB scheme becomes two-fold, since every AP-UE association attempt incorporates knowledge of APs load and mobility/load prediction is also employed.

3) Mobility-based network performance optimization approaches are extremely perceptive to mobility traces, precision of prediction and dynamic network settings. Conventionally in research, synthetically generated mobility traces are utilised for mobility modelling and prediction. In contrast, utilising real mobility traces is a demanding task due to missing values and limited context information. To the best of our knowledge, limited research efforts have been made, which optimise network resources based on real mobility traces consisting of precise location of UEs, minimal sampling time, different types of transportation modes and long-term mobility sequences. In this research work, GPS based trajectory dataset of 182 real users is used, for training and testing RNN-LSTM based mobility prediction model.

The remainder of this paper is organized as follows: Section II describes proposed MLB framework; Section III describes simulation setup and system evaluation; and Section IV concludes the paper.

II. Proposed MLB framework

The components of proposed framework (fig. 1), and their relationship with each other, are described in detail in this section. The data flow in proposed MLB framework (fig.1) is as follows:

1) Network configuration and UE measurements are captured from radio access network (RAN) in real time and are stored in database. Database comprises reported measurements of all UEs, such as UE current location, UE data requirements, UE received SINR/CQI from neighboring APs, historical mobility data, time stamped handover reports, current load of APs etc.
2) At specified time intervals, current location of all UEs and their historical mobility data is fed to mobility prediction model. This data is used as input for each prediction run and also as training data. The output of mobility prediction model is estimated coordinates for all UEs for next time instant.
3) Predicted location coordinates of all UEs are sent to load forecast module, on the basis of which future AP to UE associations are calculated. Consequently, future loads of all APs is also forecasted given the expected associations.

4) Estimated loads of all APs are sent to proactive load optimisation unit. Load optimisation unit utilises load forecasts and UE measurements from data base to estimate optimal AP to UE pairs for next time instant. These estimates are verified after performing test configurations on simulated RAN (representing deployed RAN).
5) Finally, recommendations of optimal AP-UE associations are forwarded to real RAN for implementation.

A. 5G NR RAN model

We consider a 5G New Radio (NR) UDN, comprising SCs. A single-tier network is considered where the SC APs $c = 1 \ldots C$ are distributed according to the Poisson Point Process. For simplicity only downlink (DL) communication is observed and since SCs operate in same frequency spectrum, DL co-channel interference is also assumed. The users $u = 1 \ldots U$ are randomly distributed and are served by APs through a wireless access channel. The resources scheduled to UEs are divided into fixed bandwidth Physical Resource Blocks (PRBs) and are scheduled at 1ms slot interval. The required data rate for mobile UEs is assumed to be known which gives an estimate of expected UE throughput. APs are equipped with Single Input Single Output antenna. Free Space Path Loss (FSPL) model is adopted. Logical Channels (LCHs) to support variety of applications, with logical channel prioritisation is implemented.

B. Mobility prediction model

RNNs store historical data within network's internal state and the output of RNN is dependent upon this historical input data. Generally, RNN constitutes an input layer Xt, where t denotes time instant, hidden layer(s) H and an output layer Y. In RNN-LSTM model hidden layers are substituted with LSTM blocks. LSTM blocks contain a memory cell and three gates. The memory cell controls the information flow at each time step and the gates are designed to manipulate the cell state of memory cell. As input to RNN-LSTM network, a sequence of data values $\nu 1$ to νT are given, where T is the length of sequence. The output is computed as a predicted value, $\nu T + 1$ for time $T + 1$. In the context of mobility prediction, first the position markers from the user's historical trajectory data, are extracted at a specific sampling interval. Every position marker, expressed by x-coordinate and y-coordinate, represents user's location at one specific time-step. Then a series of locations $\nu = \nu 1, \ldots, \nu T$ is formed by representing position coordinates with respect to a continuous series of time steps and this series is used as input sequence to train or test RNN-LSTM (fig. 2). Corresponding to every input sequence (user's trajectory data), the output layer maps the output of the LSTM layer at every time-step i to a position $\nu' i$, yielding output sequence $\nu' = \nu' 1, \ldots, \nu' T$, where $\nu' i$ is the predicted location at the $i + 1$ time-step [8].

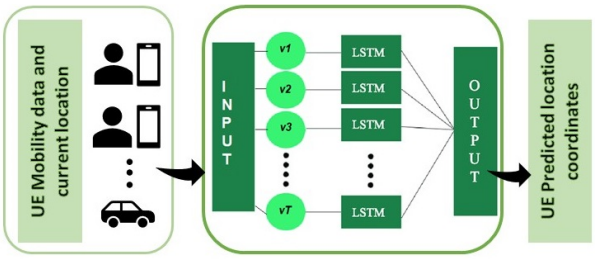

Fig. 2. Mobility prediction of network UEs.

C. Load forecast model

In the load prediction model, after predicting the possible distribution of users in next time instant, the consequent load changes in small cells is estimated. Assuming that the current time is t, then at any given instant t, future time instant can be given as $t + \Delta t$. As mentioned before $u \in U_c$, $U_c(t)$ represents all UEs associatedd with SC c at time instant t. Total number of UEs forecasted to be in c at time $t + \Delta t$ comprises of UEs who:
(i) have arrived into SC c during time interval $[t, t + \Delta t]$ and stayed in c till instant $t + \Delta t$. This set of users are given by: $U_c(\Delta t)$.
(ii) and users who were already in cell c, did not handover during $[t, t + \Delta t]$ interval and will be present in cell c at $t + \Delta t$. This set of UEs is given by: $U'_c(t)$. This set is not equivalent to $U_c(t)$ because it does not include users which handover during interval $[t, t + \Delta t]$. Therefore, the total number of UEs anticipated to be in c at $t + \Delta t$ is expressed as:

$$U_c(t + \Delta t) = U_c(\Delta t) + U'_c(t) \tag{1}$$

Cell load δ_c can be calculated as the ratio of PRBs occupied in a SC c by associated UEs, according to their required data rate and PRBs available in c. Therefore, cell load can be expressed as:

$$\delta_c = \frac{1}{N_{\mathrm{PRB}}} \left(\frac{1}{B_{\mathrm{PRB}}} \left(\sum_{U_c} \frac{Q_u}{\mathrm{SINR}_u^c} \right) \right) \tag{2}$$

where N_{PRB} is available PRBs at c, Q_u is throughput required by a UE $u \in U_c$ and U_c represents all UEs associated with SC c. SINR_u^c is SINR of u when associated with SC c and B_{PRB} is the bandwidth per PRB. Downlink SINR experienced by UE u from c is expressed as the ratio of RSRP measured by u from c to the sum of RSRP measured by u from all the interfering cells $i \in 1, ..., I$. Considering radio propagation effects, the SINR_u^c from c to u at next time instant $t + \Delta t$ can be calculated as:

$$\mathrm{SINR}_u^c(t + \Delta t) = \left[\frac{P_u^c g_{u,c}}{\sum_{i \neq c} P_u^i g_{u,i} + N_o} \right]_{[t+\Delta t]} \tag{3}$$

TABLE I
SYMBOL DEFINITIONS

Symbol	Definition	Symbol	Definition
C	Set of all SCs	$\mathrm{SINR}_{\mathrm{th}}$	Threshold SINR
$\delta_c(t+\Delta t)$	future load of c for time instant t+Δt	U	Set of all UEs
δ_c	Load of SC c	D_m	Capacity of AP m
M	Set of APs in a cluster	D_{th}	Threshold capacity
B_{PRB}	Bandwidth per PRB	δ_{um}	Load of AP m while associated with u
Q_u	Desired user throughput	$\sigma_{\mathrm{cluster}}$	Standard deviation of load of cluster
SINR_u^c	SINR at u while associated with c	σ_{th}	Threshold standard deviation of load
N_{PRB}	Total PRBs	δ_{avg}	Average load of cluster
$U_c(t+\Delta t)$	total number of UEs forecasted to be in c at $t+\Delta t$	$U_c'(t)$	users who were already in cell c and did not handover during [t, t + Δt]
$U_c(\Delta t)$	users arrived in c during [t, t + Δt] and stayed in c till instant t + Δt	$U_c(t)$	set of all UEs associated with SC c at time instant t
$\delta_c(t+\Delta t)$	load of cell c at t + Δt	$\delta_c'(t)$	load of cell c incurred by users $U_c'(t)$
$\delta_c(\Delta t)$	load of cell c incurred by users $U_c(\Delta t)$	$\delta_m(t+\Delta t)$	load of AP m at t + Δt
SINR_u^m	SINR of u when associated with m	Q_u'	Throughput delivered to u

where P_u^c is RSRP from c to u, $g_{u,c}$ is channel gain between c and u, P_u^i is transmition power of interfering cell i, $g_{u,i}$ is the channel gain between interfering cell i and u and N_o is the Additive White Gaussian Noise. The t + Δt subscript in (3) shows that all the parameters within brackets are calculated for future time instant.

Now, exploiting future user-cell association information from (1), load definition from (2) and expression to estimate SINR for next time instant from (3), the future load of small cell c for time instant t + Δt is expressed as:

$$\delta_c(t+\Delta t) = \ \delta_c(\Delta t) + \delta_c'(t) \tag{4}$$

D. Proactive load optimisation model

Within the MLB framework, the mobility prediction and load forecast module forward the estimated future coordinates of all UEs and estimated load of all the APs for instant t + Δt to proactive load optimisation unit. Using this information along with information about QoE requirements of UEs, proactive load optimization unit will estimate the optimal associations between APs and each UE. The proposed methodology is described in Algorithm 1.

1) Step 1 - Virtual cluster formation: Based on the predicted location coordinates of every UE, the APs surrounding UE on the basis of UE to APs distance are identified. From these APs, a virtual cluster of M APs such that m=1,...,M with high received SINR/CQI to UE u is created. The capacity D_m of each m in the cluster is determined and if any of the AP is already loaded according to predicted UEs distribution i.e., has surpassed its maximum capacity threshold D_{th}, it is eliminated from the cluster and cluster is recreated. Load balancing will be performed locally within each cluster of the network.

2) Step 2 - Estimating optimal Load-Aware AP-UE pairs: Instead of associating UE with an AP yielding highest received SINR/CQI in M, the predicted load of APs in M obtained using expression in (4) is also considered before association. Now for each m, algorithm

Algorithm 1 Proactive load optimisation algorithm

```
Get predicted location coordinates [x_u, y_u] of all UEs u ∈ U for next time instant t + Δt
for all u ∈ U do
    Identify surrounding APs from c ∈ C and compute their SINR_c^u(t + Δt)
    Sort SINR_c^u of each identified c to u in decreasing order
    Identify and cluster M APs with highest SINR/CQI SINR_c^u
    for all m ∈ M do
        acquire predicted load δ_m(t + Δt)
        if D_m < D_th for t + Δt is satisfied then
            Estimate δ_um if u is associated with m
            Calculate σ_cluster for δ_um for t + Δt
    Sort σ_cluster computed for each δ_um
    Find smallest σ_cluster value and update information of u and m pair which yielded smallest σ_cluster
Update association of all UEs with optimal APs at beginning of t + Δt
repeat
```

will approximate the updated load δ_{um} of each of m, if UE u is associated to it.

Then for each of u to m association case, standard deviation of load of the cluster $\sigma_{\mathrm{cluster}}$ will be computed which is expressed as:

$$\sigma_{\mathrm{cluster}} = \sqrt{\frac{\sum_{m\in M}\left(\delta_m - \delta_{\mathrm{avg}}\right)^2}{M}} \tag{5}$$

where δ_{avg} is average cluster load and δ_m is load of AP m. A smaller value of $\sigma_{\mathrm{cluster}}$ corresponds to fairer distribution of AP load within a cluster. As minimizing $\sigma_{\mathrm{cluster}}$ is the objective of LB algorithm, the optimization problem can be formulated as:

$$\begin{aligned} \underset{\delta_{\mathrm{um}}}{\text{minimize}} \quad & \sigma_{\mathrm{cluster}} \quad m \in 1,2,...,M \\ \text{subject to} \quad & C1: \sigma_{\mathrm{cluster}} < \sigma_{\mathrm{th}} \\ & C2: \mathrm{SINR}_u^m > \mathrm{SINR}_{\mathrm{th}} \\ & C3: Q_u' \geq Q_u \quad \forall u \in U \end{aligned} \tag{6}$$

TABLE II
SIMULATION PARAMETERS

Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value
DL Carrier frequency	2.635 GHz	Downlink Tx Power	32 dBm	Traffic model	Full buffer	Number of RBs	160
DL Bandwidth	30 MHz	Number of UEs	60 active UEs	Radius of cell	500 m	Packet size	8000 bytes
UL Carrier frequency	2.515 GHz	Number of SCs	15 SC APs	DL packet scheduler	Best CQI	Antenna gain Rx	10 dBi
UL Bandwidth	30 MHz	Cluster size	15 APs	Subframe duration	1 ms	Simulation time	7 days

Constraint C1 specifies that σ_{cluster} should be less than a predefined threshold value. C2 ensures that SINR should be above a threshold value. C3 ensures each user's minimum guaranteed throughput is satisfied.

Given the standard deviation values of cluster for each UE-AP pair, the association which yields least standard deviation will be chosen as optimal pair. We can summarise the load aware user association strategy with AP m as:

$$U_m := \left\{ \forall u \in U \mid m = \underset{\forall c \in C}{\operatorname{argmax}} \left(\left(\frac{1}{\sigma_{\text{cluster}}} \right)^{\alpha} \times (\text{SINR}_u^c)^{1-\alpha} \right) \right\} \tag{7}$$

where U_m indicates all UEs associated to m for which the product of the SINR and standard deviation of cluster load, is maximized for AP m. Moreover, $\alpha \in [0, 1]$ assigns weights to SINR and standard deviation of load's measurements, to manage their impact in the user association decision.

3) Step 3 - Updating AP-UE associations for next time slot: Based on suggested optimal AP-UE pairs, every UE will be connected to the respective APs at the start of t + Δt. Note that AP-UE pairs values remain fixed for one complete time interval of 1 minute, and are updated after that.

III. Simulation Results and Performance Analysis

A. Simulation setup

The 5G Toolbox and Deep Learning Toolbox in Matlab software are used, for implementation of 3GPP 5G NR specifications compliant functions and mobility prediction module [9]. 5G Toolbox™ offers the nrGNB and nrUE objects for creating the 5G base station (gNB) and user equipment (UE) nodes, respectively, for network simulation. These nodes are implemented with protocol stacks comprising the application, radio link control (RLC), medium access control (MAC), and physical (PHY) layers. The simulation parameters details are given in Table II. The prediction interval of 1 minute is considered in simulation. Therefore, every minute, our proposed MLB framework performs future location estimation of UEs for next time period and load optimization. For simulation duration of 1-day, total number of evaluation points become 1440 as per 1-minute granularity (24 hours×60 minutes = 1440 minutes). As we ran our simulation for 7 days, it yields a total of 10080 evaluation points (1440×7). At each evaluation point, performance metrics are noted to gauge prediction accuracy of mobility model and efficiency of load optimisation algorithm.

B. Network training

In our work, we used the LSTM network with 128 hidden units. The optimization algorithm used was Adam optimizer. During training of prediction model, learning rate of 0.001 and 200 epochs were used. The length of input sequences in (timesteps) in the data set is variable, depending on the length of each travel trajectory (in terms of distance). The training goal is to minimise the value of the loss function i.e. Mean Square Error (MSE).

C. Data set

The dataset utilised for training and testing the RNN-LSTM model, is comprised of traces of real vehicle trajectories. The transportation means of users in these trajectories include, car or taxi, bus and bike. These trajectories were created via GPS loggers and GPS phones installed inside vehicles. The GPS devices logged the data at a sampling rate of 1 to 5 seconds. The dataset is provided by Geolife project and is explained in detail in [10]. Each trajectory in dataset corresponds to a completed trip of a vehicle, and contains 7 features, but the features relevant to this work are latitude, longitude, altitude and timestamp.

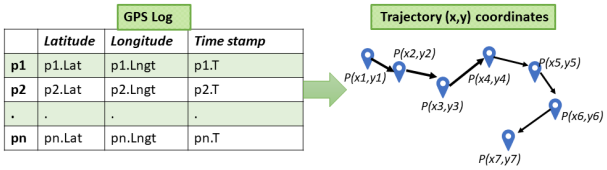

Fig. 3. Conversion of GPS logs to Cartesian coordinates.

To better suit our MLB framework we reprocessed the trajectory dataset by first converting GPS points to time-stamped cartesian coordinates, so that location of vehicles at every time stamp is represented by x-y coordinates in meters (fig. 3). Keeping in perspective, the limited coverage of SCs, we preferred trajectories within a certain range in terms of distance. Also, the vehicle trajectories which are continuous are selected. For the RNN-LSTM network, to learn the UEs mobility patterns, we used 80%

of the dataset made of UE trajectories, for the training phase, and 20% of trajectory dataset for testing phase.

D. Performance analysis

We have gauged proposed framework's performance against following; (i) The first scheme [5] predicts future load state of cells by using Bayesian Additive Regression Trees (BART) model and then balances future loads by adjusting CIO via a heuristic algorithm. (ii) The second scheme [6] formulates a Load Balanced Handover Minimized User Association (HMUA-LB) problem to minimise the number of handovers experienced by each UE and the number of overloaded APs in network, (iii) The maximum SINR based user association scenario.

To evaluate prediction performance of the RNN-LSTM model, at every evaluation point prediction accuracy is computed which is measured as the location estimation distance error (in meters) between predicted coordinates and actual coordinates of UEs. A total of 10080 average distance errors collected over 7 days simulation time, are visualised in a CDF plot in fig. 4. Minimum error was calculated to be 4 meters and maximum error was noted to be 33 meters. It is observed that majority of errors lie between 10 to 20 meters.

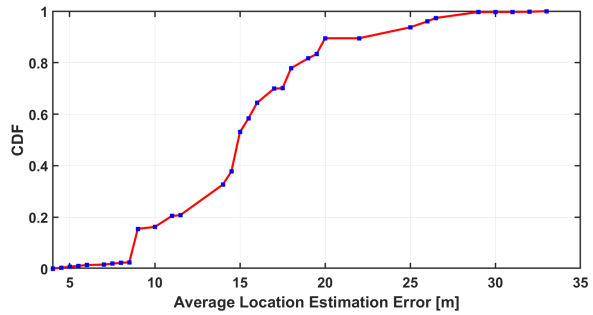

Fig. 4. The CDF of average location estimation/distance error in meters.

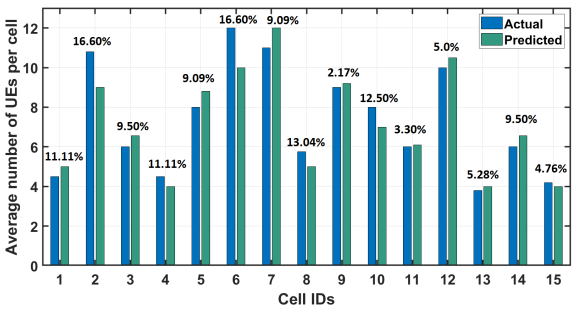

Fig. 5. Actual and predicted numbers of UEs per cell and percentage error.

A decrease in error values with increase in simulation time is observed which is attributed to the length of input sequences to prediction model. As the user starts moving and mobility trajectory becomes longer, input sequence also becomes longer, hence more knowledge is available for mobility prediction model to make a prediction. Then the correlation between actual and forecasted number of UEs in each cell is calculated. Figure 5 shows comparison of average number of actual and forecasted UEs in each of 15 cells. The percentage error is also displayed for every cell. As shown in the fig. 5, the RNN-LSTM model predicted the presence of UEs in most of the SCs with high accuracy.

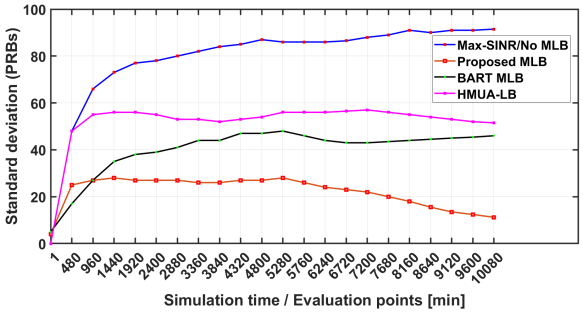

Fig. 6. Standard deviation of load comparison.

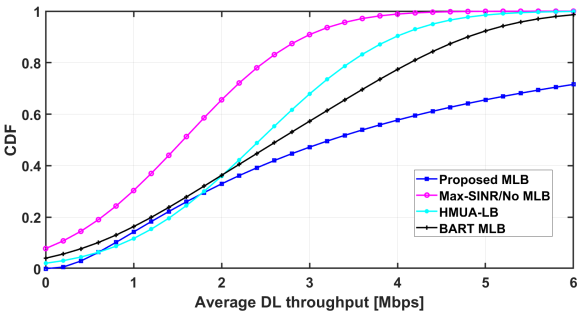

Fig. 7. Average UE DL throughput comparison.

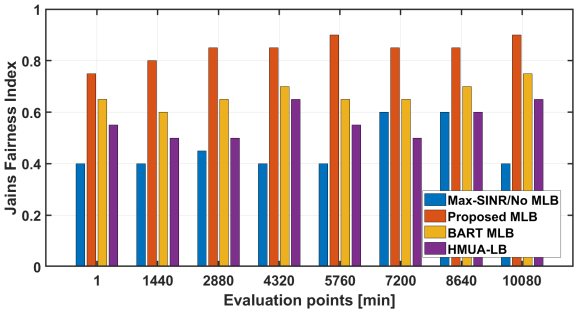

Fig. 8. Jain's Fairness Index comparison.

To evaluate efficiency of load optimisation of proposed MLB framework we used average of standard deviation of load of AP clusters VS evaluation points as our Key Performance Indicator (KPI). Smaller value of average

standard deviation at any time instant reflects better performance of algorithm and vice versa. As shown in fig. 6, maximum SINR based AP-UE association, causes worst load imbalance as load of APs is not considered. LB algorithm in [5] performs better than LB algorithm in [6] as it incorporates load forecasting. In proposed framework, future cell loads are visualised in advance, hence load and QoS aware AP selection decisions are made ahead of time and smaller deviation from average AP load is reported for all time instants. The standard deviation of load is further reduced as accuracy of mobility prediction increases due to increase in length of input sequence. This accuracy further enables the AP selection algorithm to make better selection decisions for UEs, hence better load balance.

Fig. 7 shows CDF plot of average DL UE throughput calculated across evaluation points during the simulation. The proposed MLB algorithm maintained good throughput for all UEs with maximum throughput observed to be 6.1 Mbps. This is owed to the optimal selection of AP for each UE, with least load and best SINR, which ensures that the data rate requirements of UE are satisfied. The expected drop in throughput caused by non-uniform and changing distribution of UEs, is compensated by mobility prediction assisted, proactive AP-UE association. BART MLB [5] provides comparatively better throughput than HMUA-LB [6], but still suffers with throughput degradation. This can be explained by the aggravated interference and decreased SINR for UEs which are shifted from overloaded to underloaded cells via adjusting CIO, without taking other UE QoE requirements into context.

Finally, the variance in cell loads is examined by using Jain's Fairness index plotted in fig. 8. Jain's Fairness Index β is used to express the degree of balanced distribution of load among SCs. It is understood from the fig. 8 that Fairness index for proposed MLB framework is higher than other schemes with β approaching maximum fairness of 0.9.

IV. CONCLUSION

A novel proactive, spatio-temporal mobility prediction based MLB framework for high mobility UEs in SC UDNs is presented. Extensive simulations leveraging real mobility trajectory data indicate, the proposed MLB not only maintained high downlink throughput, it also ensured balanced distribution of loads among SCs, which is critical for efficient resource utilisation in the network. Proposed MLB solution, in contrast to traditional reactive LB schemes, does not transfer load to lightly loaded cells or optimizes load after load imbalance occurs, rather it intelligently connects highly mobile users to suitable cells by utilising mobility/load prediction and load aware AP-UE association algorithm.

REFERENCES

[1] C. Wang, Z. Zhao, Q. Sun, and H. Zhang, "Deep learning-based intelligent dual connectivity for mobility management in dense network," 2018.

[2] Y. Ma, X. Chen, and L. Zhang, "Base station handover based on user trajectory prediction in 5g networks," in *2021 IEEE Intl Conf on Parallel Distributed Processing with Applications, Big Data Cloud Computing, Sustainable Computing Communications, Social Computing Networking (ISPA/BDCloud/SocialCom/SustainCom)*, 2021, pp. 1476–1482.

[3] J. Luo, Y. Chen, J. Hu, P. Chen, and H. Zheng, "Intelligent load balancing relying on load prediction with mgcn-gru," in *2022 IEEE/CIC International Conference on Communications in China (ICCC)*, 2022, pp. 884–889.

[4] H. Farooq, A. Asghar, and A. S. Imran, "Mobility prediction based proactive dynamic network orchestration for load balancing with qos constraint (opera)," *IEEE Transactions on Vehicular Technology*, vol. 69, pp. 3370–3383, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:213222227

[5] M. Huang and J. Chen, "Joint load balancing and spatial-temporal prediction optimization for ultra-dense network," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, 2022, pp. 506–511.

[6] S. Biswas, A. Gupta, and S. Chakraborty, "Load-balanced user associations in dense lte networks," *Computer Networks*, vol. 189, p. 107928, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128621000761

[7] M. Shabbir, S. Kandeepan, A. Al-Hourani, and W. Rowe, "Access point selection in small cell ultra-dense network, with load balancing," in *2022 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, 2022, pp. 1–6.

[8] H. Zhang, Y. Hua, C. Wang, R. Li, and Z. Zhifeng, *Deep Learning Based Traffic and Mobility Prediction*, 12 2019, pp. 119–136.

[9] T. M. Inc., "5g toolbox," 2023. [Online]. Available: https://au.mathworks.com/help/5g

[10] Y. Zheng, X. Xie, and W.-Y. Ma, "Understanding mobility based on gps data," in *Proceedings of the 10th ACM conference on Ubiquitous Computing (Ubicomp 2008)*, September 2008. [Online]. Available: https://www.microsoft.com/en-us/research/publication/understanding-mobility-based-on-gps-data/