Telecommunication Systems
https://doi.org/10.1007/s11235-024-01202-w

Optimizing cell association in 5G and beyond networks: a modified load-aware biased technique

Mohammed Jaber Alam[1] · Ritesh Chugh[2] · Salahuddin Azad[2] · Md Rahat Hossain[1]

Accepted: 16 July 2024

Abstract
Cellular networks are moving towards increasing heterogeneity by deploying more small cells into macro base station (MBS) to meet rapidly growing traffic demands. To leverage the advantages of these small cells, mobile users should be offloaded onto small base stations (BSs), which will typically be lightly populated and can give a higher data rate by presenting the mobile users with many more channels than the MBS. Likewise, a more balanced cell association will lessen the pressure on the MBS, allowing it to serve its remaining users more effectively. This paper addresses the cell association challenge for Quality of Service (QoS) provisioning in terms of throughput and load-balancing for 5G and future generation networks. This problem is quite challenging because BSs have varying backhaul capacities and users have varying QoS needs. Most of the previous studies are based on reference signal received power (RSRP), signal to interference and noise ratio (SINR) or its variants and most importantly majority of them are not load-aware. Therefore, a modified load-aware biased cell association scheme based on distance is proposed to attain better QoS provisioning in terms of throughput and load-balancing. Simulation results depict that the proposed load-aware-based method outperforms conventional cell association schemes based on RSRP and its variants, and in terms of throughput and load-balancing. Furthermore, the algorithm's complexity has been assessed through a comparison and analysis of computational time, demonstrating better performance compared to state-of-the-art techniques.

Keywords 5G and beyond · Heterogeneous network · Load balancing · Cell association · Ant colony optimization

Abbreviations

BS	Base station
QoS	Quality of Service
RSRP	Reference signal received power
SINR	Signal to interference-plus-noise ratio
HetNet	Heterogeneous network
MBS	Macro base station
IoT	Internet of Things
UE	User equipment
CRE	Cell range extension
AP	Access point
MIMO	Multiple-input multiple-output
SBS	Small base station
3GPP	3rd generation partnership project
BPP	Binomial point process
LA	Location area
LM	Location management
LU	Location uppdate
GSM	Global system for mobile communications
BAR	Biasing association region

✉ Mohammed Jaber Alam
m.alam@cqu.edu.au

Ritesh Chugh
r.chugh@cqu.edu.au

Salahuddin Azad
s.azad@cqu.edu.au

Md Rahat Hossain
m.hossain@cqu.edu.au

[1] School of Engineering and Technology, CQUniversity, North Rockhampton, QLD 4701, Australia

[2] School of Engineering and Technology, CQUniversity, Melbourne, VIC 3000, Australia

1 Introduction

Heterogeneous Networks (HetNets) are now considered a significant architecture of fifth generation (5G) and future generation networks since they have been proven to be a promising paradigm for increasing data rates and user

Published online: 18 August 2024

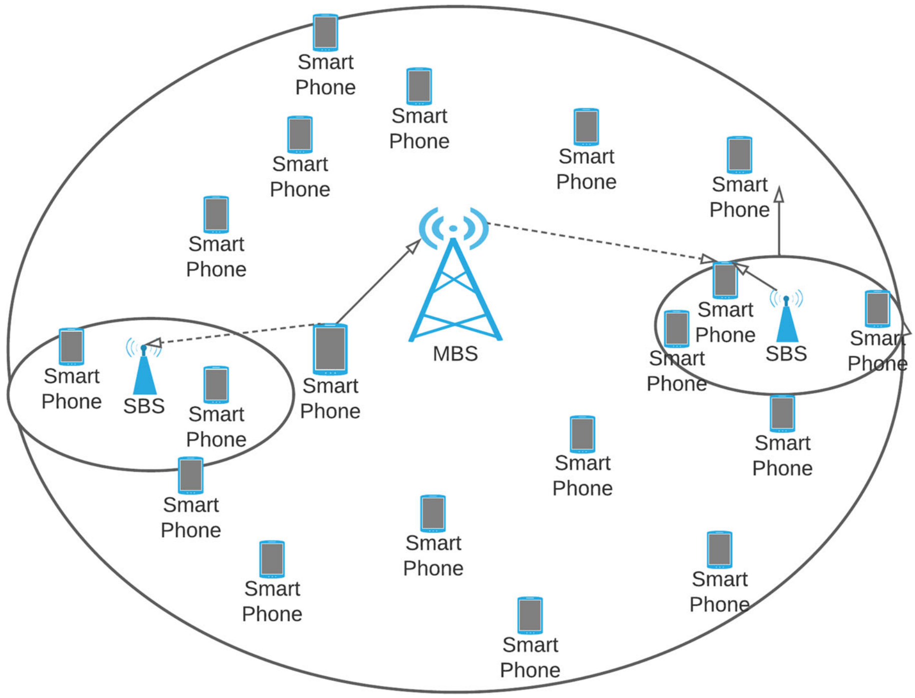

Fig. 1 Heterogeneous network architecture

capacity. In HetNets, small cells, also referred to as small base stations (SBSs), are installed within macro base station (MBS) to enhance the coverage areas which are inadequately served (See Fig. 1 for an illustration of a heterogeneous network) [1]. The major differences between 4G HetNets and 5G HetNets are higher data rates, lower latency, increased capacity, enhanced coverage and intelligent network management, among many others. 5G HetNets enable a wide range of new use cases and applications that were not feasible or practical with 4G networks. These include Ultra-Reliable Low-Latency Communications (URLLC), massive Machine-Type Communications (mMTC), and mission-critical applications. Addressing the challenges of enhancing the Quality of Service (QoS), load balance, throughput, and interference mitigation in 5G HetNets are thus more critical. Therefore, many studies have been ongoing to address the problems and meet the standards of 5G and future generation networks [2].

Heterogeneous networks have been a topic of interest because of their role and benefits in establishing Internet of Things (IoT) and smart cities. Researchers explored this field mainly from a resource allocation perspective and introduced novel resource allocation schemes for heterogeneous networks with an overarching aim to help multimedia applications [3]. On the other hand, some other researchers delved into the technical challenges from a load-balancing perspective [4]. This research direction has turned out to be increasingly significant because user association, which is a key load-balancing mechanism, is generally executed before resource allocation. Hence, it significantly affects the network performance. Therefore, inefficient cell association causes load imbalance and wasteful resource distribution among User Equipment (UE), limiting QoS provisioning [5] [6].

One key strategy for advancing 5G success involves implementing HetNets, incorporating Base Stations (BSs) with high transmission power like MBS alongside those with lower transmission power, such as PBSs and Femto Cells. The primary benefits of deploying HetNets include boosting network capacity and enhancing link quality for users who are in closer proximity to BSs [7]. Nevertheless, conventional cellular user association methods tend to result in the majority of users associating with MBS, drawn by their higher transmission power, even when smaller cells like pico or Femto Cells are closer. This conventional association approach leads to a load imbalance, with MBS experiencing overload while SBSs have lighter loads [8]. Along with this, the throughput of cell-edge users of the MBS drops, which often leads to call-drop and cell-edge user being out of service. In addition, lot of other users associated with the MBS also suffer from low data rate leading to lower average network throughput. Thus, it is a struggle not only for users to get the best service but also for the operators maintaining the claim of 5G standards [9]. Despite the existing research on load-balancing for heterogeneous networks, there have not been adequate robust solutions to meet the 5G and beyond user demands, which motivates this study.

This paper is structured as follows. The next section outlines the literature review, followed by the system model in Section III. Next, in Section IV, the proposed algorithm illustrates the proposed load-aware-based biased cell association scheme. Section V outlines the results and discussion. Finally, a conclusion is presented in Section VI.

2 Literature review

In existing research works such as [10], the load carried by each BS is stated as the number of UEs served, or the quantity of resources consumed. Although, backhaul capacity, which is generally the constriction for the load that each BS can handle, is not taken into account. Therefore, the conventional load-balancing technique in [11] is not directly applicable to the practical load-balancing of base stations. Load-balancing of base stations is co-related to the backhaul capacity of the base station. Each base station has its own backhaul capacity; hence, the load-balancing problem arises because of the disparity of backhaul capacity of SBSs and MBS in a heterogeneous network. This leads to a disproportionate distribution of users among base stations. Although [12] addressed load-balancing, the different QoS user requirements have not been analyzed.

Previous research on load-balancing techniques relates mainly to networks consisting of only MBS [13]. However, HetNets comprising of SBSs (such as PBSs and Femto Cells) and MBS are more sensitive to the cell association strategy because of the significant variance in cell sizes. These disparities result in much more unbalanced loads in a signal to interference and noise ratio (SINR) or reference signal received power (RSRP) based cell association, where the mobile user distribution has been assumed to be uniform. So, if the users simply link up with the strongest base station, the difference of load in MBS networks is constrained since all these base stations have equal coverage areas. Therefore, recently some advanced RSRP methods, such as RSRP+biased association and RSRP+power control association have been introduced [14]. RSRP+power control is the method of blanking MBS power at certain instances to allow its associated users to be shifted to the neighboring small base station. This power control allows a better load balance than the general RSRP method. Meanwhile, RSRP+biased association method introduces an additional bias to the RSRP. This bias is generally added to the small base station power, to combat higher MBS transmission power disparity. This method is also named Cell Range Extension (CRE).

The authors in [15] address the crucial role of the hand-off process in WiFi networks, particularly in scenarios with numerous Access Points (APs), where even a few seconds of delay can lead to information loss in time-sensitive applications. The paper, therefore, claims to achieve better throughput with reduced packet loss. In [16], an overview of load-balancing along with cell association techniques is mentioned, which proclaims most state of art techniques revolve around RSRP and SINR. In [17], the authors conducted a comparative examination of various user association strategies within HetNets, incorporating biased user association. The paper presents AURA-5G and demonstrates that the optimal solutions enhance the performance of diverse network scenarios in terms of total network throughput and system fairness when compared to the conventional cell association scenario.

Cell range extension [18], enabled through appropriate cell biasing, is an effective method to balance the load among macro and small base stations. It is obtained by employing adaptive cell association based on the biased signal strength. This aids in better load-balancing; however, this often degrades the bit rate (throughput) that certain users receive. Hence, one of the critical open problems is how to design a suitable biasing factor [19]. Yasin Aghazadeh et al. proposed a method in [20] that uses estimated received signal strength from different cells and adjusted pilot signals. However, the authors did not include the least visited Femto Cell access points, which probably limits the efficacy of the technique by getting unnecessary hand-offs. They also did not discuss about the complexity and scalability of the proposed technique.

Researchers are exploring the potential advantages of incorporating Cell Range Expansion (CRE) to improve the performance of heterogeneous networks. Numerous studies, including [24] and [21], propose adjusting the CRE area based on the SINR, while others advocate expanding the region through the analysis and updates of the co-ordinated scheduler operation, as outlined in [21]. To enhance HetNet performance, Cell-Selection Offsets (CSOs) can be employed to modify the CRE's coverage area, as suggested in [25]. An innovative concept presented in [4] involves dynamically adjusting both the CRE region and Almost Blank Sub-frame (ABS) based on the traffic conditions of each cell in the network [9]. Furthermore, researchers have explored various techniques to enhance the user experience in 5G systems, such as interference coordination in massive MIMO (Multiple-Input Multiple-Output) networks during the transition from 4G to 5G [26]. Table 1 illustrates a qualitative comparison of cell association techniques with our proposed technique.

The current study investigates load-balancing of base stations by considering the load of MBS and Small Base Stations (SBSs) and therefore also realising the QoS requirements of users. This paper emphasizes on downlink load-balancing among base stations, along with throughput maximization for 5G and beyond. Therefore, in this paper, a load-aware-based cell association method based on the distance from heterogeneous downlink networks is proposed. The proposed algorithm will allow UEs to associate with the base stations by distance rather than conventional RSRP, which will improve the load balance of the network along with the QoS (in terms of user throughput or bit rate). The study is novel

Table 1 Qualitative comparison of cell association techniques specifically for load-balancing

Technique	Key focus	Methodology	Limitations
CRE [4]	QoS enhancement	CSO and SINR-based adjustments	Interference mitigation is not considered
CRE and TD-ICIC [14]	QoS enhancement	Base stations are expanded and muted	Optimal results are not yielded
Conventional cell association [18]	Throughput	RSRP	Throughput drops for cell-edge users
Cell Range Extension (CRE) [19]	Data rate	RSRP+biased	Interference increases, hence optimal throughput not achieved
CRE [20]	Throughput	RSRP+power control	Increased interference
Proportional Fairness (PF) resource scheduling-based adaptive algorithm [21]	CSO optimization	Adjusted bias value via PF scheduling	Resource utilization efficiency
NP-hard based network-wide cell association [22]	Load balancing	max-SINR	Does not evaluate either interference or throughput simultaneously in the results
Energy-efficient load aware cell association [23]	Energy efficiency and load balance	Energy-Efficient Load-Aware User Association (EELUA)	UE mobility is not considered during simulation
Proposed (load-aware-based))	Load-balancing and throughput	Dynamically adjusting bias for each BS based on load-awareness	Trial and error method to choose the bias

because we can see from the literature that the state-of-the-art techniques are mainly based on RSRP and SINR and rarely consider load-awareness. But as HetNets are proceeding towards ultra-dense networks, a new proposition such as load-aware distance based cell association could be a solution. The simulation results from our experiments in this work add merit to our claim.

3 System model

The simulation has been performed in MATLAB software following the specifications of Release 16, 3GPP TS 24.426 version 16.5.0. We have considered a downlink network containing one MBS overlaid by 3 SBSs. Table 2 depicts the summary of the main parameters used in the simulation. The network can be conceptualized as a multi-tier cellular network, where base stations belonging to the same tier share identical transmission power, density, and coverage area. Furthermore, all tiers are autonomously distributed using the Binomial Point Process (BPP) with a density λ_k for tier k. The BPP model assumes that BSs are deployed randomly and independently of each other, meaning that the location of each BS is chosen independently of the locations of other BSs. This assumption closely resembles real-world deployment scenarios where BS locations are determined without consideration for the locations of neighboring BSs.

Table 2 Parameters used in the simulation

No. of MBS	1
No. of PBS	3
O	refers to the BSs
X	refers to the UEs
No. of users	500
PBS radius	50 m
MBS radius	500 m
Each sub-channel bandwidth	180 KHz
Total No. of sub-channels,K	100
Transmit power of MBS	43 dBm
Transmit power of SBS	30 dBm
Channel	Rayleigh fading
Noise power Spectral density	– 174 dBm/Hz
Antenna gain	5 dB
User noise figure	9 dB
PBS path loss model	140.7 + 36.7 * log10(d) dB (d [Km])
MBS path loss model	128.1 + 36.7 * log10(d) dB (d [Km])
Shadowing standard deviation	10 dB
No. of trials	1000

By assuming random and independent BS deployment, the BPP model provides a simplified yet realistic representation of the spatial distribution of BSs in heterogeneous networks. This allows researchers and engineers to analyze network performance, coverage, and interference in a mathematically tractable manner while capturing key characteristics of real-world deployment scenarios. The reason most researchers use BPP during system modelling is its scalability feature, meaning that it can be applied to networks of varying sizes without significantly increasing computational complexity. For the sake of simplicity in notation, we designate the BS subset as $S = 1, 2, \ldots, S$, with the first element representing the MBS. Each BS j is characterized by the tuple P_j, B_j, χ_j, wherein P_j signifies transmit power in equation (1). Assuming independent and identically distributed Rayleigh fading between BS j and user location x, denoted as $h_j(x)$, the received power at user location x is given by $P_j h_j(x)(d_j)^{-\alpha_j}(x)$, where $h_j(x) \sim \exp(1)$, α_j is the path loss exponent, and $d_j(x)$ is the distance from user location x to the BS j. Consequently, the received SINR of BS j at user location x can be articulated as,

$$\gamma_j(x) = \frac{P_j h_j(x)(d_j)^{-\alpha_j}(x)}{I(x) + N_0} \tag{1}$$

Here, $I(x)$ represents the interference emanating from both the identical tier and diverse tier BSs at the user location x, while N_0 stands for the noise power spectral density. Utilizing the provided $\gamma_j(x)$, we employ the Shannon capacity to characterize the transmit rate in the following manner:

$$R_j(x) = W \log_2(1 + \gamma_j(x)) \tag{2}$$

For the above, W denotes the bandwidth of base station j.

4 Proposed load-aware biased cell association scheme

Conventionally, the users are associated with the base stations by the RSRP. The more the transmission power of a cell, the more the probability that the user gets associated with it. However, there are path losses, fading and interference that hamper the RSRP signal received by a user. The MBS, being the bigger base station, generally has more users connected to it, while the small base stations, such as PBSs do not get that many users associated with them. This is because the users are associated with the base station with higher signal strength. The load-balancing problem becomes more challenging when each base station in the heterogeneous networks has a different backhaul capacity. For example, MBS has a higher backhaul capacity than PBS. This disparity of BS capacities underutilizes the small base stations by a disproportionate distribution of users among the base stations, creating an imbalance among the load shared by the base stations. Moreover, the MBS's overloaded users do not get enough resource blocks or bandwidth required for 5G and beyond standards.

Algorithm 1 Proposed algorithm

```
Initialize the network parameters such as path-loss, antenna gain, sub-channel bandwidth, transmit power of BSs, etc.
Initialize all variables to zeros :

Each UE u measures the SINR based on the signal from each BS s, and estimates R_su and γ_su by equations (1) and (2), respectively.

Each UE u sends the information of R_su and γ_su to each BS s.

Repeat:

The user locations are updated to the BS register.

The distances of all the users u from each base station s are calculated.

Users are associated with the nearest base stations based on distance calculation.

The load of each BS is checked.

IF:

Load on one BS more than the threshold, the furthest cell-edge users are handed over to the next nearest BS.

Bias, β is added to the SBS so that the cell-edge user of MBS can connect to the closest SBS.

Users are associated by the distance plus the bias of user from the particular BS.

Else:

Cell association remains same as previous.

End:

Re-association is done and throughput is calculated.

Throughput and Jain's Fairness Index (JFI) are calculated accordingly.

Repeat the above process for X times for an instance and choose the best throughput and JFI.
```

In this paper, a load-aware based cell association is proposed, as shown in Algorithm 1, where the users connect with the base stations with regards to distance from users to base stations, rather than conventional RSRP. In this scheme, the closer a user is to a BS, the higher the probability of that user being connected to the base station. Therefore, the disparity of power capacities of MBS and SBSs is not a factor anymore. Also, if the percentage of the total number of users connected to a BS reaches a threshold, the furthest cell edge users of that BS is handed over to the nearest BS. The percentage can be set by the operator (in this case by us), and it is understood that at peak hours and non-peak hours, the rush of users can vary. A bias is added to the nearest SBS so that the handover occurs accordingly from the MBS to the SBS, whenever the threshold percentage is crossed for the MBS. The bias value is also decided based on the distance of user being handed over to the nearest BS, the further the user is, the bigger the bias value is needed. The first step of this scheme is to get the location of the user equipment. The BSs are fixed, and so their location coordinates are known. In effect, the more precisely we can locate the users, the more precisely we can calculate the distance of the users from these base stations.

4.1 Location management methodology

Location Areas (LAs) are groups of network cells that users move across, updating the network with their location as per set standards. When a user receives a call, the network pages the cells in the LA to locate the user, a process known as Location Management (LM). The network can reduce paging costs by requiring frequent Location Updates (LUs), but this increases time and energy costs. Conversely, infrequent LUs lower these costs but increase paging costs. LAs can be optimized to reduce hand-offs and speed up location updates. The goal of LM is to balance these factors. Most current LM systems are static, with LUs happening periodically or with each cell change. However, frequent updates occur when users repeatedly switch between two or more LAs, known as the ping-pong effect. Static LAs are the same for all users, uniform in size and fixed. The current static LM standards, GSM (Global System for Mobile Communications) and IS-41 (also known as ANSI-41) have a hierarchical database structure [27]. LUs are executed via one of these criteria:

1. Update location with each BS transition: A location update is initiated whenever a UE relocates to a different BS.
2. Page all cells within the network: When an incoming call needs to be directed to a UE, a page is broadcasted to all cells in the network to determine the cell associated with the UE.
3. Divide the network into paging sub-regions: As a UE moves to a new paging sub-region, it informs the network about the identity of that sub-region. When an incoming call is received for that UE, only the cells within the current sub-region are searched to identify the cell linked to the UE. This entails additional criteria such as optimizing the location update process with reduced traffic load and refining the paging procedure to minimize traffic congestion.

The distance between users and BS is calculated from their respective geographic coordinates, enabling the association of UE with the nearest BS. This methodology incorporates a load-aware mechanism, ensuring that if a specific BS becomes overloaded-defined by surpassing a predetermined threshold of users-the system dynamically reassigns the furthest user to the next nearest BS. This intelligent handover mechanism alleviates congestion and optimizes network performance by distributing the user load more evenly across available BSs. Following the user association phase, resource allocation is executed using conventional techniques such as round-robin scheduling, wherein resource blocks are allocated uniformly among users. This uniformity in resource characteristics ensures the allocation process is consistent across all scenarios examined in this study. By leveraging these strategies, the proposed approach not only aligns with established network management practices but also incorporates advanced load-balancing techniques to enhance efficiency and service quality. This comprehensive method of user association and resource allocation highlights the robustness and scalability of the proposed network management framework in 5G and beyond HetNets. The resource allocation has been carried out similarly for all the schemes discussed in this paper.

4.2 Methodology of adding bias

Primarily, we assume unrestricted access for all base stations, allowing users to connect to any cell within the network coverage area. To balance the load among BSs, we implement a scheme based on distance plus the bias in case of hand off required. In this scheme, users selectively connect to the BS with the highest biased distance. Consequently, the biased distance can be expressed as follows:

$$\gamma_j(x) = \gamma_j(x) + B_j \tag{3}$$

where B_j is the biasing factor of BS j. Note that the biasing factor of the MBS is always set to be $B_1 = 1$, which is the same as the traditional max-SINR cell association. Generally, set $B_j > 1$ for $j \geq 2$ to make it more attractive for users. Unlike the static biased scheme, we design our proposed load-aware based scheme to achieve global load-balancing. Considering the non-uniformly distributed traffic, the biasing factor changes along with the load of BS and the traffic arrival rate. Intuitively, the SBSs located in high traffic density areas may reduce the biasing factor to shrink coverage, while the SBSs located in low traffic density areas may increase the biasing factor to attract more users.

Next, we present the definition of the Biasing Association Region (BAR) of BS j. Users located within the coverage will associate with that particular BS j based on the maximum biased SINR. Thus, the covering radius with biasing can be expressed as

$$r_{b,j} = B_j^{1/\alpha_j} \cdot r_j \tag{4}$$

where r_j the covering radius without biasing can be obtained from the equation below,

$$r_j = \left(\frac{P_j g_j(x)}{\chi_j (I(x) + N_0)} \right)^{1/\alpha_j} \tag{5}$$

Here, P_j is the transmit power for the BS j, $g_j(x)$ is the channel gain of BS j from user x, $I(x)+N_0$ is the addition of interference and noise power respectively. To further clarify, χ_j represents the received SINR threshold and α_j is the path-loss exponent of BS j.

5 Results and discussion

An MBS of 500 m radius overlaid with 3 PBSs located is considered, as shown in Fig. 2. The MBS is positioned at the center, and the PBSs are placed at random locations, as indicated in Fig. 2. The power transmitted by the PBSs and MBS has been set to 30 dBm and 43dBm, respectively. 500 UEs are considered, which are randomly distributed within the MBS and SBSs, which is again BPP distribution, and their required data rate is randomly set. The key factors to calculate the data rate are the path loss models, transmit power, bandwidth, noise power, and Rayleigh fading. Table 2 shows the values of these parameters used in this study.

To model the overall probability distribution of data rates, the following steps are followed:

1. Simulate user placement: Randomly place 500 users within the MBS and SBS coverage areas.
2. Assign data rates: Calculate the data rate for each user based on their distance to the nearest base station and the path loss model.
3. Fit distribution: Fit a suitable distribution (e.g., log-normal) to the calculated data rates.

The log-normal distribution is used to model data rates in this context, as follows:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2} \right) \tag{6}$$

where μ and σ are parameters estimated from the simulation data. This equation represents the Probability Density Function (PDF) of a log-normal distribution. The variables are defined as follows:

- $f(x)$: The probability density function of the log-normal distribution evaluated at x.
- x: The data rate (in this context), which is a positive real number. It represents the variable for which we are calculating the probability density.
- μ: The mean of the natural logarithm of the data rates. It is a parameter of the log-normal distribution estimated from the simulation data.
- σ: The standard deviation of the natural logarithm of the data rates. It is another parameter of the log-normal distribution estimated from the simulation data.
- $\ln x$: The natural logarithm of x.
- exp: The exponential function.
- $\sqrt{2\pi}$: The square root of 2π, a normalizing factor in the PDF of the log-normal distribution.

From the calculation, the data rate range for MBS is 95 Mbps-1 Gbps, and SBS is 51 Mbps-491 Mbps. BPP assumes that users are distributed randomly and independently of each other, which closely resembles real-world deployment scenarios in wireless networks. The required data for each user is also random, but each channel the user uses has limited bandwidth. The channel contains 100 Physical Resource Blocks (PRBs), with each having 180 kHz bandwidth. The path loss models used for PBSs are 140.7 + 36.7 log(d) (dB) and for MBS is 128.1 + 36.7 log(d) (dB), where d is the distance measured in kilometers (kms) between the BS and the UE. The noise spectral density and noise figure are set to -174 dBm/Hz and 9dB, respectively. Zero-mean log-normal shadowing with 10dB standard deviation and a channel with zero-mean unit-variance Rayleigh fading is considered. In addition, the proposed algorithm (distance based load-aware biased cell association) is compared with the maximum RSRP, RSRP+biased, RSRP+power control, max-SINR and EELUA cell association schemes with the same resource allocation mechanism. The simulation results for each scheme are run for 1000 instances, and then the average reading is considered to combat the random nature of the simulation parameters like interference, path-loss and fading. This is important because, in practical scenarios, these parameters are similarly random in nature, and this consideration in our simulation makes the comparison of the results stronger. In each instance, users have been stationary for all the compared schemes as well as the proposed one but located at various random positions to consider the mobility and experiencing different channel conditions.

In Fig. 3, the network's average throughput [13], is evaluated against different number of users. The proposed load-aware-based association scheme achieves significantly higher throughput because the power disparity is avoided during user association. The power disparity in the RSRP method pushes more users to be associated with MBS rather than SBSs, which results in the throughput being shared among

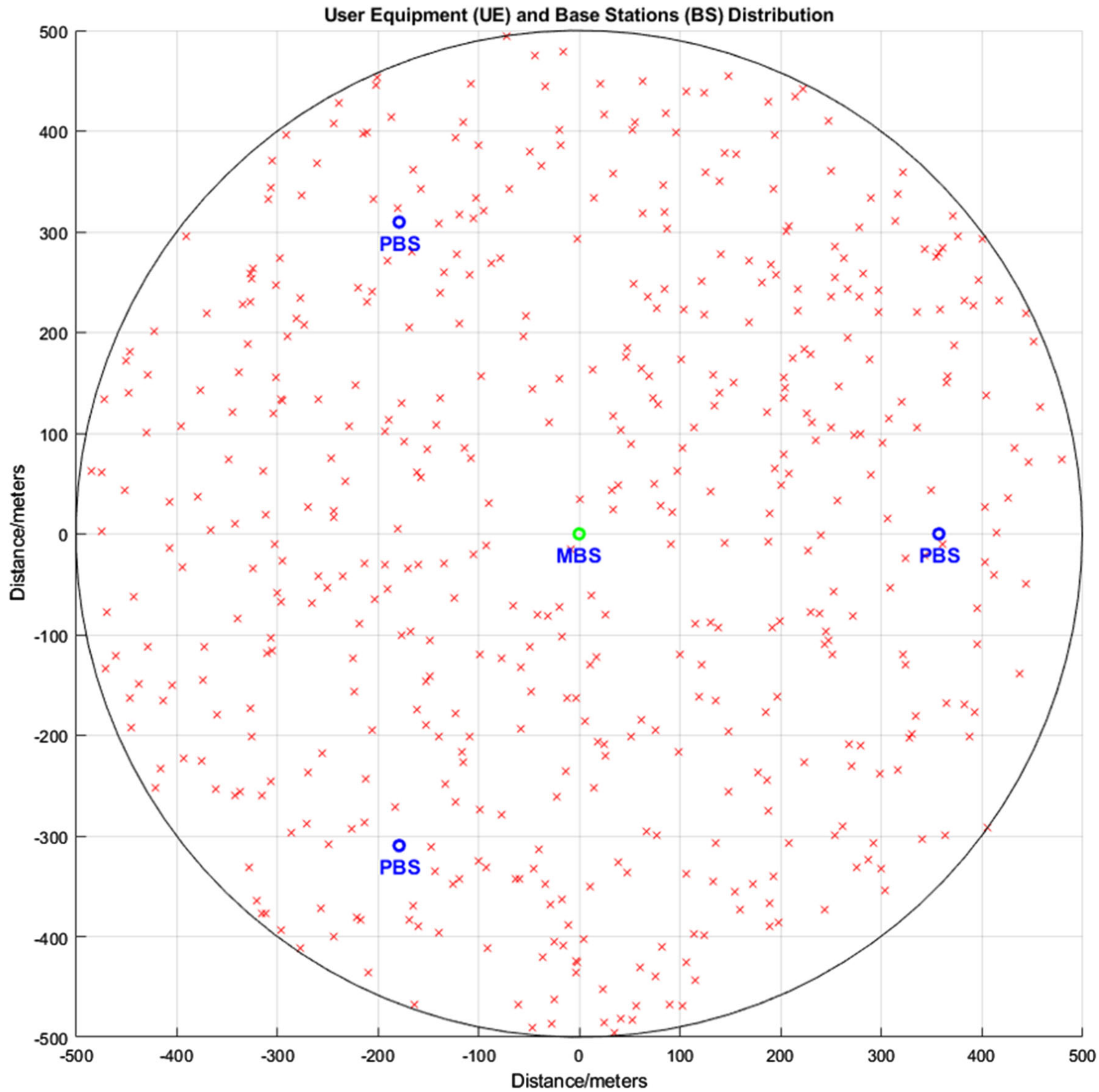

Fig. 2 Simulation of system model

a large number of users in the congested MBS, leading to a relatively low throughput per user. As most users are associated with MBS rather than SBSs in the RSRP scheme, the average throughput of the network drops significantly, unlike the proposed distance-based load-aware scheme. Even the other methods presented in Fig. 3, (e.g., RSRP+biased and RSRP+power [14] control) could not sufficiently enhance the throughput in all circumstances (different number of users). The max-SINR based association method as well as EELUA method performs poorly as the number of users increased.

Load-balancing has been depicted by Jain's fairness Index (JFI) [13] in Fig. 3, which is implemented to evaluate the fairness of distribution of users among BSs or the load-balancing [18] of the heterogeneous network. It is generally the best way to compare and indicate the efficiency of load distribution per cell versus different numbers of users. The proposed technique is found to outperform the other methods though it follows a similar downward trend as the user number increases, like the other compared schemes. However, it can be observed from Fig. 3 that when the number of users is 300, the throughput performance drops while JFI peaked. This might be due to a sub-optimal bias value. However, the overall throughput and JFI of the proposed technique outperform the other techniques in most instances. The proposed distance-based load-aware scheme provides 10–12% more load balance compared to the other network schemes. The downward trend for a higher number of users might be because of the UEs being randomly distributed (random walk) in the simulations, the number of users associated

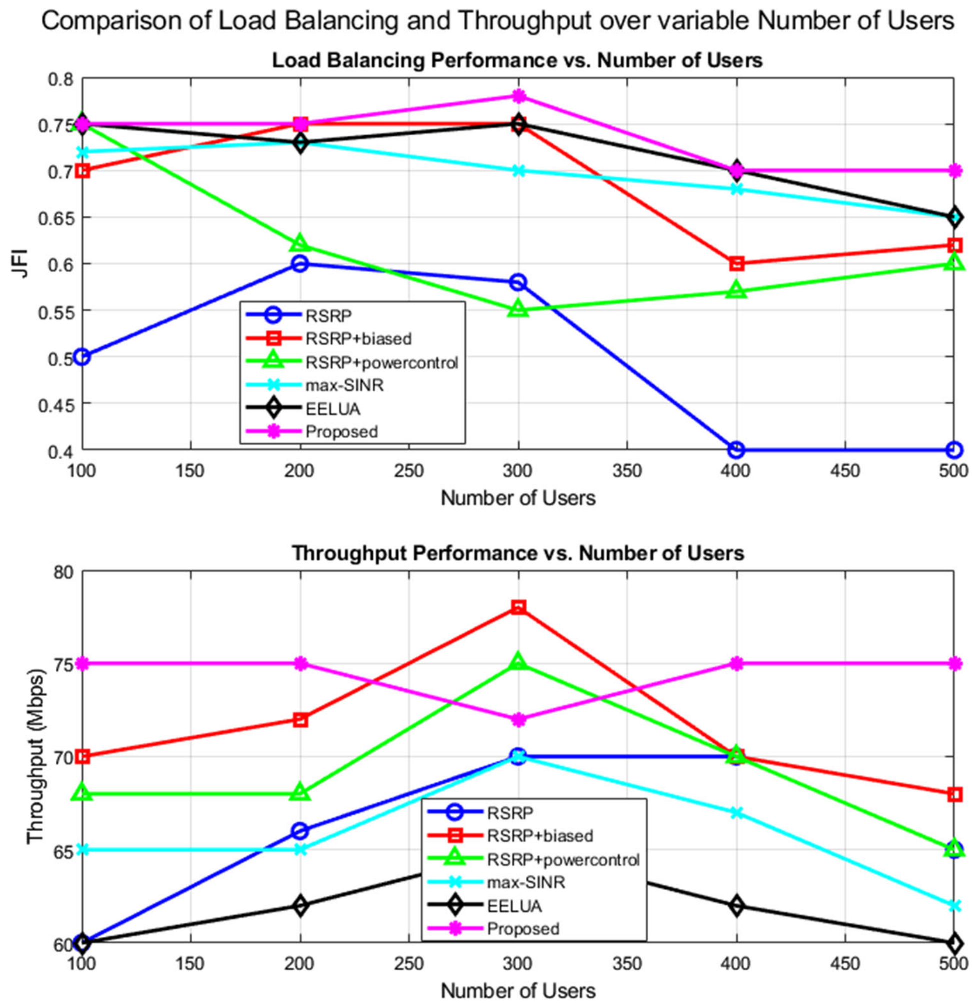

Fig. 3 Performance comparison of proposed scheme against the state-of-the-art techniques over variable number of users

with some of the base stations increases, leading to a lower throughput per user.

Furthermore, the throughput and JFI is evaluated against variable hand-off numbers in Fig. 4. These findings provide valuable insights into how both the proposed methods and current techniques perform under conditions where users are moving at high speeds, which is an important aspect to consider in real-world scenarios. The trend of load balancing and throughput among BSs versus hand-off number in a HetNet can vary depending on various factors such as network topology, traffic patterns, and hand-off management strategies. It can be observed from Fig. 4 that the proposed technique outperforms the other state-of-the-art techniques in almost all hand-off conditions. While techniques like EELUA or RSRP+biased or RSRP+power control lost load balance with higher hand-off number, the proposed technique has shown stability. Although in the case of throughput performance, the rate dropped for the proposed technique with the increase of hand-offs, but still sufficiently better than the other techniques. Factors such as interference from neighboring cells, signaling overhead associated with hand-offs, and variations in network conditions can influence the throughput-hand-off relationship. Therefore, while there may be variations in the specific trend depending on network configuration and user mobility patterns, the overall trend often reflects an initial improvement followed by stabilization or decline in throughput as hand-off frequency increases.

It is imperative to assess whether the advancements in throughput and load-balancing come at the expense of heightened computational complexity. Implementing a sophisticated algorithm with potential delays, particularly in the context of cell association, poses practical challenges. Thus, we meticulously compared and analyzed the computational complexity, as illustrated in Fig. 5. Notably, the proposed algorithm demonstrates a significantly shorter duration for user association and handovers between cells compared to the RSRP+biased, RSRP+power control, max-SINR and EELUA algorithms. While the load-aware algorithm requires slightly more time than the conventional RSRP technique, this trade-off is justified by the substantial improvements in throughput and load-balancing. It is easily understandable that a flat RSRP cell association does not consume as

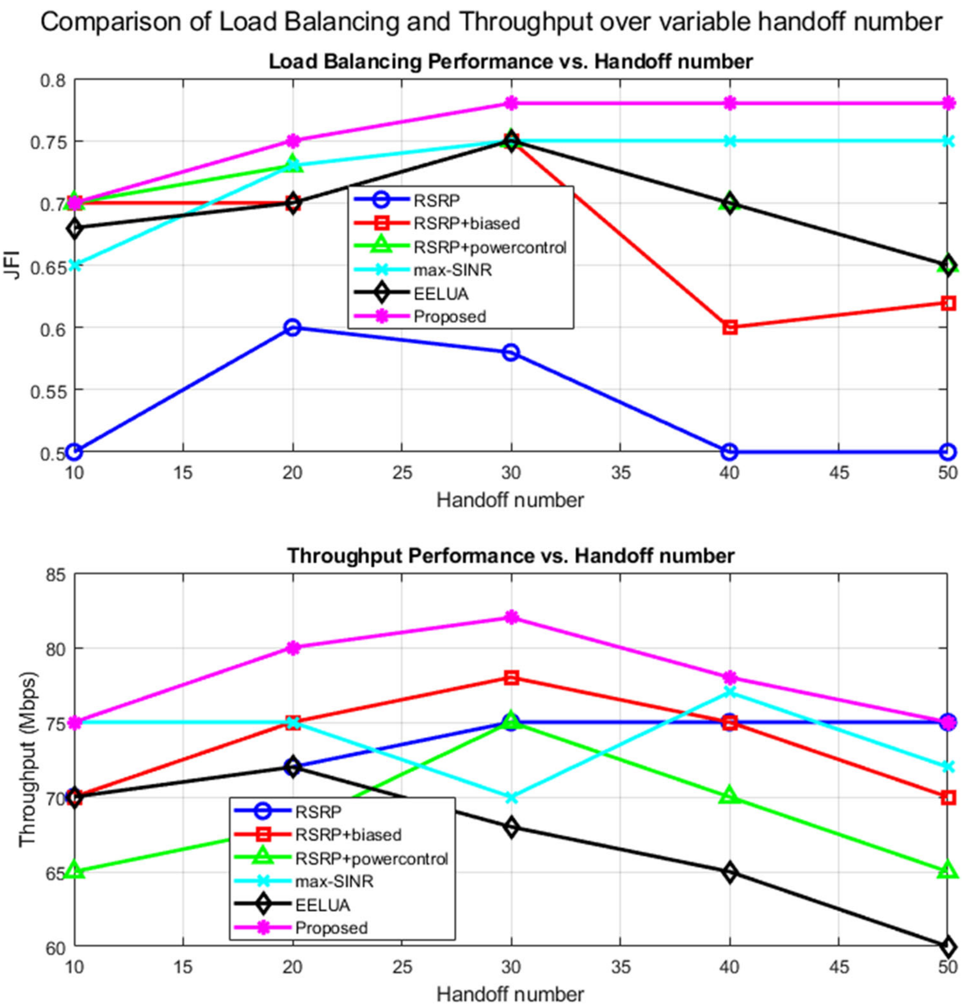

Fig. 4 Performance comparison of the proposed scheme with state-of-the-art techniques against varying hand-off numbers

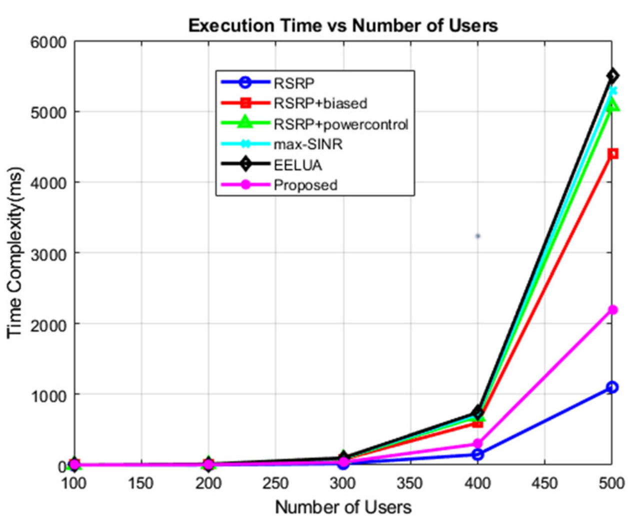

Fig. 5 Complexity analysis

much computational time compared to the other techniques which are manipulating various parameters to obtain the necessary gain. We contend that this level of computational complexity is justified, considering the substantial enhancements achieved in the targeted parameters.

6 Conclusion

This study introduces an innovative cell association approach founded on load-awareness, where the association is determined based on proximity rather than the traditional RSRP and its variants. From the state of art, it could be observed that, there is a study gap of load-aware cell association, which simultaneously solves the load-balancing and throughput challenges of 5G and future networks. Therefore, the primary objective of this proposed technique is to enhance the load distribution among base stations while concurrently ensuring QoS provisioning, specifically in terms of throughput, for 5G and future networks. The simulation outcomes demonstrate the superior performance of the load-aware-based technique over existing cell association methodologies, showcasing notable improvements in both throughput and load-balancing.

Nevertheless, it is crucial to acknowledge certain limitations, such as the use of a random user deployment model to simulate mobility, which warrants further investigation for more accurate representations of real-world scenarios. We also believe that in future work, the complexity or the computational time can be reduced by the introduction of proper optimization techniques. Future research endeavors could delve into refining user mobility modeling, and additional considerations might encompass interference mitigation, radio resource management, and bandwidth optimization within heterogeneous network environments.

Acknowledgements This study has received support from the Commonwealth Research Training Program (RTP) and CQUniversity Australia.

Author Contributions Mohammed Jaber Alam is the main contributor to this manuscript. He wrote the manuscript, carried out simulations, and undertook analysis of the results. A/Prof Ritesh Chugh revised the manuscript and provided invaluable guidance in the documentation. Dr Salahuddin Azad and Dr Md Rahat Hossain have reviewed and revised the manuscript as well as validated the results. All authors have read and approved the final manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data Availability Data sharing is not applicable to this article.

Declarations

Ethics approval and consent to participate Not applicable

Conflict of interest The authors have no conflict of interest relevant to this paper.

References

1. Hossen, A.A., Mungur, A., Armoogum, S., & Pudaruth, S. (2021). An analytical review of load balancing in heterogeneous networks. In *Proceedings of the Future Technologies Conference*, pp. 225–236. Springer.
2. Alam, M.J., Gafur, A., Rashid, S.Z., Sadeque, M., Kundu, D., Sayed, R., et al. (2022). Permutation based load balancing technique for long term evolution advanced heterogeneous networks. *International Journal of Electrical & Computer Engineering* (2088–8708) **12**(6).
3. Sheikhidris, K., Alias, M.Y., Roslee, M., & Alam, M.J. (2018). Throughput maximization based on user association in heterogeneous networks. In *2018 6th International Conference on Information and Communication Technology (ICoICT)*, pp. 34–37. https://doi.org/10.1109/ICoICT.2018.8528793
4. Chang, B. J., & Liou, S. H. (2017). Adaptive cooperative communication for maximizing reliability and reward in ultra-dense small cells lte-a toward 5g cellular networking. *Computer Networks, 115*, 16–28.
5. Hong, E.-K., Lee, I., Shim, B., Ko, Y.-C., Kim, S.-H., Pack, S., Lee, K., Kim, S., Kim, J.-H., & Shin, Y. (2022). 6g r&d vision: Requirements and candidate technologies. *Journal of Communications and Networks, 24*(2), 232–245.
6. Eguizábal, M., & Hernández, Á. (2016). Joint dynamic resource allocation and load balancing-cell selection in lte-a hetnet scenarios based on type 1 inband relay deployments. *Computer Networks, 100*, 90–109.
7. Alam, M.J., El-Saleh, A.A., Tan, C.K., Ku, I., Lee, Y.L., & Chuah, T.C. (2018). Improved joint cell association and interference mitigation for lte-a heterogeneous networks. In *2018 IEEE 4th International Symposium on Telecommunication Technologies (ISTT)*, 1–4. https://doi.org/10.1109/ISTT.2018.8701715.
8. Alam, M.J., Hossain, M.R., Chugh, R., & Azad, S. (2023). Distance-based cell range extension and almost blank sub-frame for load balancing and interference mitigation in 5g and beyond heterogeneous networks. *Engineering Reports*, 12772.
9. Alam, M.J., Hossain, M.R., Chugh, R., & Azad, S. (2023). Multi-objective solution of reducing call drop rate and maximizing throughput by ant colony optimization in 5g heterogeneous networks. In *2023 IEEE International Conference on Networking, Sensing and Control (ICNSC)*, vol. 1, 1–6. https://doi.org/10.1109/ICNSC58704.2023.10319017
10. Storck, C.R., Efrem, E.D.O., Guilherme, G.D.O., Mini, R.A., & Duarte-Figueiredo, F. (2021). Fivh: A solution of inter-v-cell handover decision for connected vehicles in ultra-dense 5g networks. *Vehicular Communications*. **28**, 100307.
11. Hasan, M. M., & Kwon, S. (2019). Cluster-based load balancing algorithm for ultra-dense heterogeneous networks. *IEEE Access, 8*, 2153–2162.
12. Chen, Y., Li, J., Lin, Z., Mao, G., & Vucetic, B. (2016). User association with unequal user priorities in heterogeneous cellular networks. *IEEE Transactions on Vehicular Technology, 65*(9), 7374–7388.
13. Zohar, N. (2022). Beyond 5g: Reducing the handover rate for high mobility communications. *Journal of Communications and Networks*. **24**(2), 154–165. https://doi.org/10.23919/JCN.2022.000001.
14. Gadam, M. A., Ahmed, M. A., Ng, C. K., Nordin, N. K., Sali, A., & Hashim, F. (2016). Review of adaptive cell selection techniques in LTE-advanced heterogeneous networks. *Journal of Computer Networks and Communications, 2016*, 3.
15. Manzoor, H., Manzoor, S., Ali, N., Sajid, M., Menhas, M. I., & Hei, X. (2021). An SDN-based technique for reducing handoff times in WiFi networks. *International Journal of Communication Systems, 34*(16), 4955.
16. Alam, M.J., Hossain, M.R., Azad, S., & Chugh, R. (2023). An overview of lte/lte-a heterogeneous networks for 5g and beyond. *Transactions on Emerging Telecommunications Technologies*, 4806.
17. Jain, A., Lopez-Aguilera, E., & Demirkol, I. (2021). User association and resource allocation in 5g (aura-5g): A joint optimization framework. *Computer Networks, 192*, 108063.

18. Rakotomanana, E., & Gagnon, F. (2016). Optimum biasing for cell load balancing under QoS and interference management in HetNets. *IEEE Access, 4*, 5196–5208.
19. Kudo, T., & Ohtsuki, T. (2013). Cell range expansion using distributed q-learning in heterogeneous networks. *EURASIP Journal on Wireless Communications and Networking, 2013*(1), 61.
20. Aghazadeh, Y., Kalbkhani, H., Shayesteh, M. G., & Solouk, V. (2018). Cell selection for load balancing in heterogeneous networks. *Wireless Personal Communications, 101*, 305–323.
21. Gadam, M. A., Ng, C. K., Nordin, N. K., Sali, A., & Hashim, F. (2016). Hybrid channel gain prioritized access-aware cell association with interference mitigation in LTE-advanced HetNets. *International Journal of Communication Systems, 29*(18), 2650–2668.
22. Ye, Q., Rong, B., Chen, Y., Al-Shalash, M., Caramanis, C., & Andrews, J. G. (2013). User association for load balancing in heterogeneous cellular networks. *IEEE Transactions on Wireless Communications, 12*(6), 2706–2716.
23. Mughees, A., Tahir, M., Sheikh, M.A., & Ahad, A. (2021). Energy-efficient load-aware user association in ultra-dense wireless network. In *2021 26th IEEE Asia-Pacific Conference on Communications (APCC)*, pp. 254–259. IEEE.
24. Alam, M. J., Chugh, R., Azad, S., & Hossain, M. R. (2024). Ant colony optimization-based solution to optimize load balancing and throughput for 5g and beyond heterogeneous networks. *EURASIP Journal on Wireless Communications and Networking, 2024*(1), 44.
25. De Domenico, A., Díez, L. F., Agüero, R., Kténas, D., & Savin, V. (2015). Emf-aware cell selection in heterogeneous cellular networks. *IEEE Communications Letters, 19*(2), 271–274.
26. Giambene, G. (2018). Analysis of LTE-a heterogeneous networks with sir-based cell association and stochastic geometry. *Journal of Communications and Networks, 20*(2), 129–143.
27. Mukherjee, A., & De, D. (2016). Location management in mobile network: A survey. *Computer Science Review, 19*, 1–14.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Mohammed Jaber Alam has received his Bachelor of Science degree in Electronics and Telecommunications Engineering in 2016. He passed his Master of Engineering Science by research degree in 2019 from Multimedia University, Malaysia. Simultaneously, he was a Research Scholar at the same institution for a couple of years. He is at the end of completing his PhD from CQUniversity, Australia. He is also a Research Fellow and a Casual Academic at the School of Engineering and Technology, CQUniversity, Australia. He previously worked as a Research Fellow at the ICT Ministry in Bangladesh. Jaber is an Associate Member of the Australian Computer Society as well as IEEE Communication Society. He has authored and co-authored several articles in the field of Telecommunications Engineering. Jaber's research interests include wireless and mobile communications, electronics, power, antenna and wave propagation, and artificial intelligence.

Ritesh Chugh is an Associate Professor at Central Queensland University, Australia's School of Engineering and Technology. As an information systems sociotechnological expert, his research focuses on the social role of information systems and their influence on humans and organizations. He takes an interdisciplinary research approach in areas such as information systems management, educational systems, and technology-enhanced learning, with his research featured in leading academic journals. He is acclaimed for nurturing student outcomes through reflective teaching. Notable accolades include the International Education Association of Australia's Best Practice in International Education Award and the Dean's Award for Mid-Career Research Excellence. He has also been recognized with the Vice-Chancellor's Award for Good Practice in Learning and Teaching and a Staff Award for International Education contributions. In mainstream media, Ritesh engages in public scholarship on sociotech topics, nationally and internationally. He is a Senior Fellow of the Higher Education Academy, further affirming his dedication to advancing educational excellence. Ritesh is a senior member of the Institute of Electrical and Electronics Engineers and the Australian Computer Society.

Salahuddin Azad is an experienced academic in the field of Computing and Information Technology. Dr Azad received the B.Sc. Eng. (hons.) degree in Computer Science and Engineering from Bangladesh University of Engineering and Technology (BUET) in 1999 and the PhD degree in Information Technology from Monash University, Australia in 2007. He is currently working as a Lecturer in the School of Engineering and Technology at Central Queensland University (CQU). He has 20 years of track record of teaching in Australia and overseas. He previously worked as a postdoctoral research fellow at the Queensland University of Technology (QUT). His research interests include wireless communications, machine learning, natural language processing, and image processing. He is serving an expert reviewer for Australian Research Council (ARC) College.

Md Rahat Hossain is a Senior Lecturer at the School of Engineering and Technology of CQUniversity, Australia. He is a dedicated and articulate academic with extensive teaching and learning experience in Information and Communications Technology (ICT). He possesses a record of ability to conduct high-quality research, which is reflected in the authorship of high-quality publications, and other research outputs, in the areas of ICT, Renewable Energy, and Regional Studies. He has been supervising students at the Masters and PhD levels and has a few research scholarships and grant awards. Rahat has authored and co-authored numerous articles in the field of Engineering and is dedicated to the research and academic field.