Computers, Materials & Continua
DOI:10.32604/cmc.2021.015490
Article

Tech Science Press

Intelligent Real-Time IoT Traffic Steering in 5G Edge Networks

Sa Math[1], Prohim Tam[1] and Seokhoon Kim[2,*]

[1]Department of Software Convergence, Soonchunhyang University, Asan, 31538, Korea
[2]Department of Computer Software Engineering, Soonchunhyang University, Asan, 31538, Korea
[*]Corresponding Author: Seokhoon Kim. Email: seokhoon@sch.ac.kr
Received: 24 November 2020; Accepted: 09 January 2021

Abstract: In the Next Generation Radio Networks (NGRN), there will be extreme massive connectivity with the Heterogeneous Internet of Things (HetIoT) devices. The millimeter-Wave (mmWave) communications will become a potential core technology to increase the capacity of Radio Networks (RN) and enable Multiple-Input and Multiple-Output (MIMO) of Radio Remote Head (RRH) technology. However, the challenging key issues in unfair radio resource handling remain unsolved when massive requests are occurring concurrently. The imbalance of resource utilization is one of the main issues occurs when there is overloaded connectivity to the closest RRH receiving exceeding requests. To handle this issue effectively, Machine Learning (ML) algorithm plays an important role to tackle the requests of massive IoT devices to RRH with its obvious capacity conditions. This paper proposed a dynamic RRH gateways steering based on a lightweight supervised learning algorithm, namely K-Nearest Neighbor (KNN), to improve the communication Quality of Service (QoS) in real-time IoT networks. KNN supervises the model to classify and recommend the user's requests to optimal RRHs which preserves higher power. The experimental dataset was generated by using computer software and the simulation results illustrated a remarkable outperformance of the proposed scheme over the conventional methods in terms of multiple significant QoS parameters, including communication reliability, latency, and throughput.

Keywords: Machine learning; Internet of Things; traffic steering; mobile edge computing

1 Introduction

Fifth Generation (5G) communication utilizes millimeter-Wave (mmWave) technology to deliver ultra-high communication reliability and throughput, maximum user devices' connectivity, heterogeneous user applications, etc. [1]. The 5G Stand-Alone (SA) based communication systems have fully been migrated from the legacy Long-Term Evolution (LTE) system. In the SA-based radio networks, the Next Generation evolved Node B (gNB) provides controllability throughout the Control Plane (CP). For Non-Stand-Alone (NSA) radio networks, LTE's Radio Remote Head (RRH) utilizes both evolved Node B (eNB) and gNB. It is worth mentioning that the NSA is generally suffered from insufficient slicing control of the application Quality of Service (QoS) and inadequate software-based radio resource control. Since the NSA 5G system has been widely launched in several countries, and there are serval remaining challenging issues to fully migrate to the SA communication systems. The 5G Radio Access Network (5GRAN) communication systems will be converged by numerous computing power devices and recent edge technologies. So, the 5GRAN environments will upsurge potential power with incredible resources and opportunities to overcome the issues of massive connectivity for the Internet of Things (IoT) devices. Due to the explosive growth of mobile devices and heterogeneous IoT (HetIoT) devices, big data has been rapidly generated throughout the network. The huge user traffic volume has suffered from insufficient capacity at the fronthaul and backhaul gateways which are commonly installed by optical network environments. To minimize outgoing traffic to the remote cloud, namely the Mobile Cloud Computing (MCC), the mobile edge cloud technology is a vital key candidate for local computing perspectives.

Furthermore, 5GRAN architecture consists of massive infrastructures and platforms that enable local cloud servers for handling heterogeneous services of HetIoT and Mobile Users (MU) (see Fig. 1). The increment of Ultra-Dense Networks (UDN) and mesh connectivity of wireless networks are offered by numerous base stations, such as macrocell, microcell, picocell, femtocell, massive MIMO, etc. According to the extensive growth of small cell Base Stations (BS), a lower transmission power is continually applied for small cell coverage purposes. Microcell, picocell, femtocell, relay, and Device-to-Device (D2D) offer short-range distance communications with lower or equal power to 30 dBm, and from 0 to 5 dBi for the antenna power gains [2]. The restriction of users in joining the radio networks occurs in the low power BS systems for both wide-band and narrow-band communications [3]. To satisfy the massive MIMO's QoS provision, the presence of software-based radio network control is mandatory. 5G enhances broadband wireless networks with high user data rates from 100 Mbps up to 20 Gbps for MU with extremely high peak rates accordingly [4]. The evolution of both mobility (wireless) and stationary technologies, including Wireless Local Area Network (WLAN), Wi-Fi, Fiber to the Home (FTTH), Passive Optical Network (PON), and Active Optical Network (AON) required to enumerate in 5GRAN communication systems [5].

To achieve dynamic efficient service provisioning for IoT and MU, the revolution and evolution of the intelligent and powerful network architecture need to be performed. Due to the fact that Over the Top (OTT) applications have rapidly increased concurrently, especially in the period of the Covid-19, the usage of Internet service and applications has exponentially increased. There are several types of Critical Communication Applications (CCA) including video conversation, voice conversation, remote operation, non-streaming conversation, game streaming, real-time IoT traffic, and other time-sensitive communication applications which needs to meet an End-to-End (E2E) latency requirement close to 0 s and 99.99% communication reliability [6,7]. To meet the requirement of CCA, E2E Ultra-Low Latency (ULL) services, Ultra-High Communication Reliability (UHCR), and Ultra-High Mobility (UHM) have to be considered. Software-Defined Networking (SDN), Mobile Edge Computing (MEC), and Network Function Virtualization (NFV) were also applied to offer superb contributions for distinct CCA requirements [8,9].

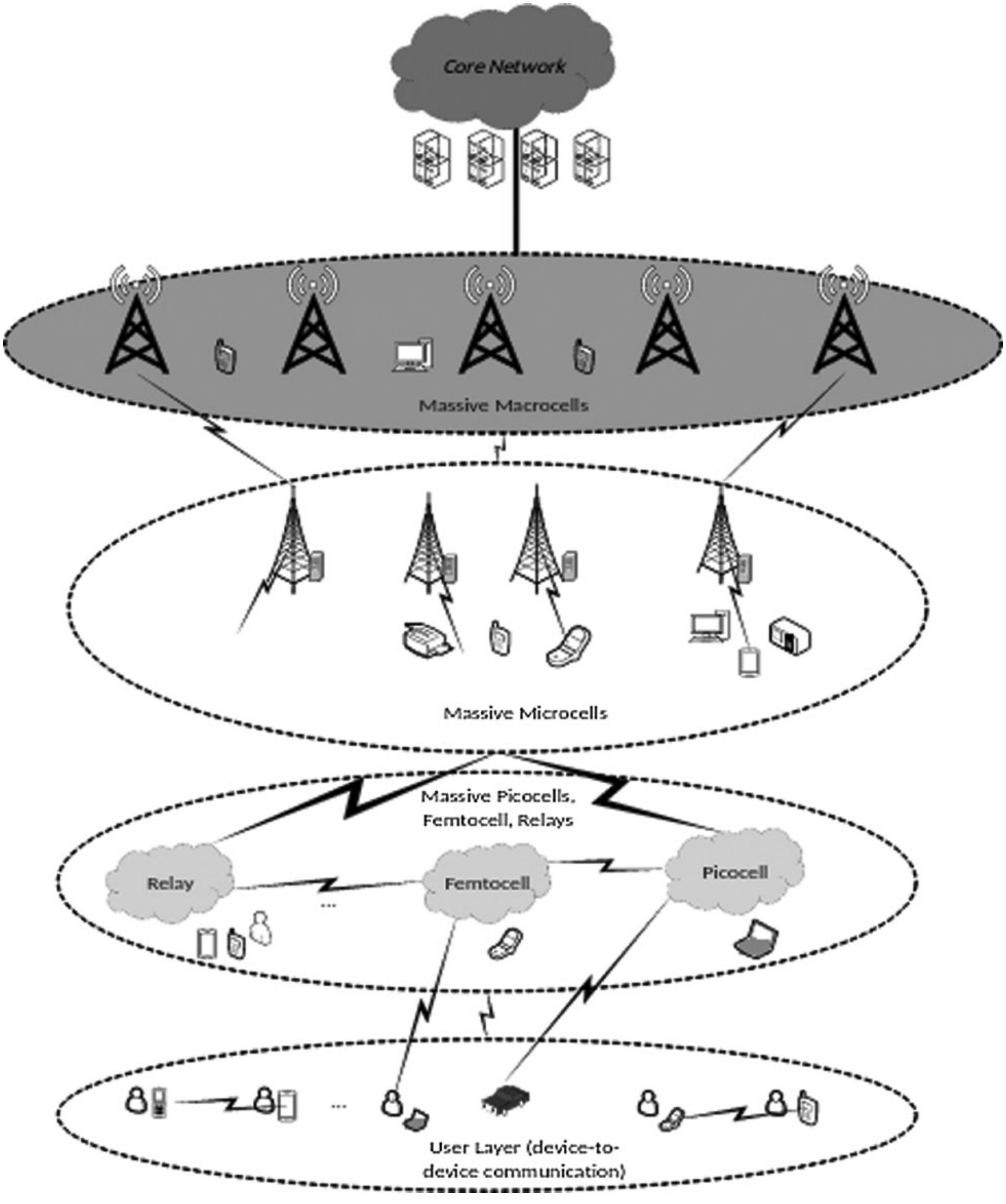

Core Network
Massive Macrocells
Massive Microcells
Massive Picocells, Femtocell, Relays
Relay
Femtocell
Picocell
...
...
User Layer (device-to-device communication)

Figure 1: The common architecture of next-generation massive 5G radio networks

The rest of the paper is organized as follows. In Section 2, the overview of the radio network environment is illustrated. The proposed traffic handling method based on a lightweight machine learning algorithm is presented in Section 3. In Section 4, the discussion is thoroughly studied about the proposed communication system analysis. In Section 5, the simulated results and analysis are intuitively described and evaluated. Finally, the paper conclusion is provided in Section 6.

2 Radio Network Environments

In this section, we provided several illustrations of significant elements for the 5GRAN environments including the involved technologies, challenging issues appearance in terms of resource management and orchestration, radio resource handling, service offloading, and opportunity enabling to overcome the significant issues mentioned. A detailed discussion of each element is presented as follows.

2.1 Technology Adoption

The 5GRAN technology arises out of heterogeneous cloud platforms, and will be a convergence of various novels and up-to-date technologies to alleviate from a legacy Radio Access Network (RAN) to a powerful radio cloud systems. MEC paradigm becomes a compulsory candidate, widespread numerous opportunities in the edge network environments, caching technology enabler, and local cloud servers [10]. The caching method extracts popular request content of applications from the remote cloud to compute in Edge Cloud Networking (ECN). The QoS can be enhanced by reducing the communication distance so that its delay can be omitted from each network device. However, the MEC-based radio cloud suffers from the expanding cloud infrastructure for both Capacity Expenditure (CAPEX) and Operation Expenditure (OPEX) [11,12]. The CAPEX is influential to the management, orchestration and the 5GRAN architecture turns into more complications. Due to the emergence of CAPEX, OPEX is increasing simultaneously, the Infrastructure as a Service, Platform as a Service, Resource as a Service, and Management and Orchestration as a Service will happen in the future 5GRAN environment [12–14].

NFV provides Virtual Network Infrastructures (VNI) which runs on top of host operating systems in single computing physical hardware. Virtual Edge Clouds (VEC) offered by NFV and provides the benefit of less cost deployment and prevents the suffering of CAPEX and OPEX [15–17]. The implementation of NFV with MEC enables VEC and also virtual 5GRAN (v5GRAN). The v5GRAN entails a variety of Virtual Machines (VM) for independent computing. The VEC synchronizes with each RRH for multiple computing purposes in terms of user request caching information, user QoS/Quality of Experience (QoE) control, Radio Link Control (RLC), uplink and downlink control, latency, and antenna gains monitoring [18]. Due to the limitation of computing power and resources of MEC, virtual MEC (vMEC) plays an important role to fulfill insufficient resources. Based on the VEC environment, the backup resources are possible for fault-tolerance offloading. Fault-tolerance boosts communication reliability and can be analogously computed for load-balancing offloading. NFV is recommended to apply at the CP area since the VM can be suffered from the offloading time, which possibly degrades the communication performance of the User Plane (UP). The 5G UP networks handle massive traffic forwarding that required wide-band and ULL for E2E communications. Meanwhile, the CP requires a redundancy computing server, high computing power and stability, fault-tolerance and load-balancing for tremendous computing services. MEC and NFV offer both physical and virtualized resources for enabling E2E Network Slicing (NS) methods. NS classifies the differences of user terminal behaviors, application, QoS/QoE, radio resource, services, etc., so the VM is required for each slicing.

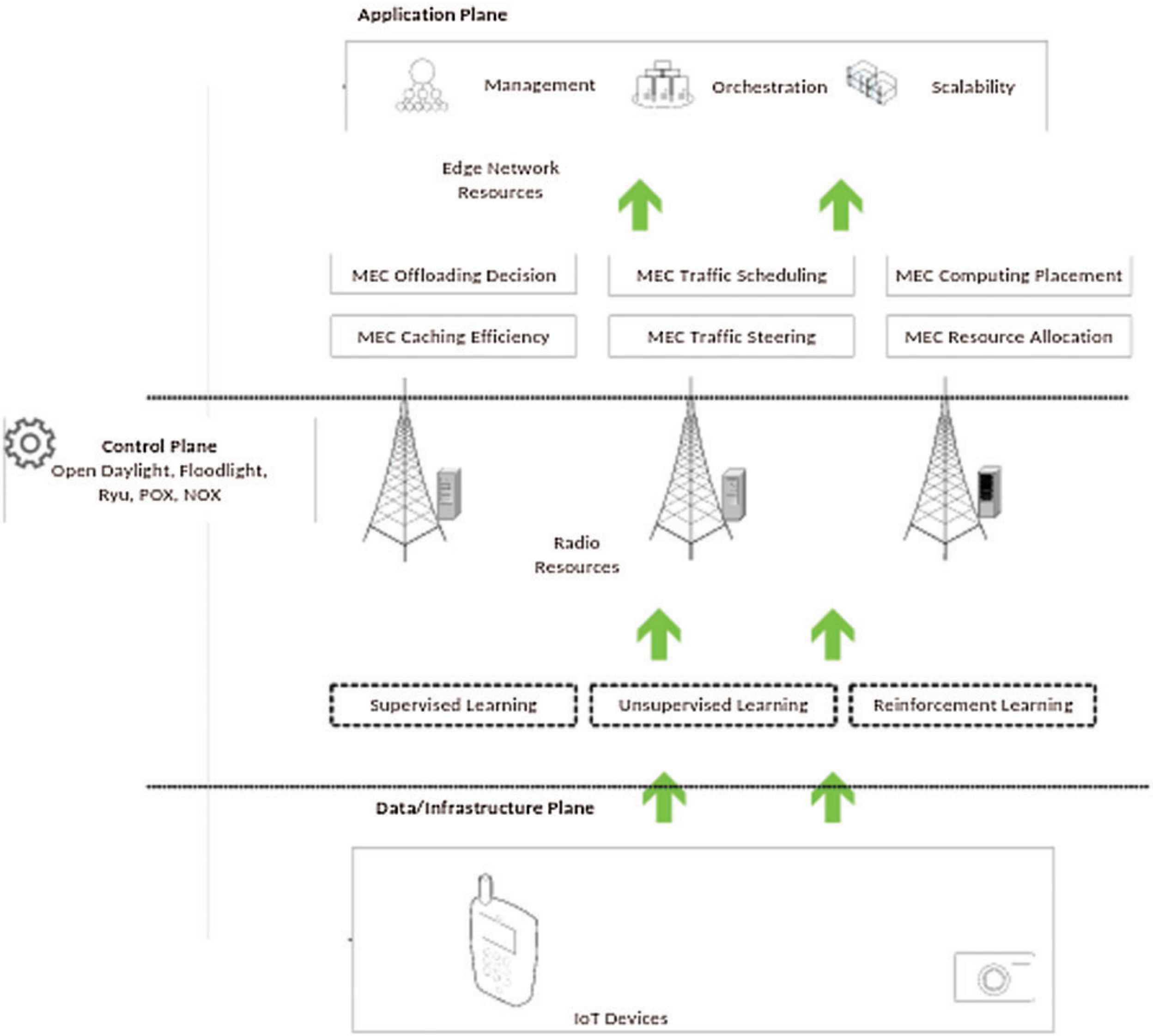


Figure 2: The convergence of ML, MEC, and SDN based architecture for next-generation radio network

Furthermore, it requires the decoupling of CP from the UP for more independent operation based on different responsibilities. The convergence of key technologies enables intelligent 5GRAN network architecture (see Fig. 2). SDN performs a decoupling of CP from the UP functions and offers application programming interfaces (API) for management and orchestration of the three layers in common SDN architecture, such as UP, CP, and application plane layers [19]. The UP, sometimes called the data plane, takes the responsibility of forwarding the incoming traffic to reach the destination based on the configured flows table offered by SDN controller. The ULL forwarding devices with wide-bandwidth are capable of forwarding massive user data at the UP. The UP is a forwarding plane area without computing or making decisions, and the forwarding flow configuration executes at the CP functions. The communication flows between southbound and northbound interfaces are linked based on the OpenFlow protocol [20]. The CP offers the computing infrastructure for a variety of services. SDN controller has a global view of the network systems due to its location in the centralized application and data plane infrastructures. So, the data plane device's capacities or conditions are monitored by the controller module. SDN takes important roles in the management and orchestration of the multiple VM in the VNI. The integration of SDN and NFV enables numerous opportunities since both SDN and NFV fulfill the missing points to each other for better performances [21,22]. 5GRAN architecture is beneficial from the combination of SDN/NFV. The existing entities of the legacy radio system can be fully replaced by SDN and NFV in both UP and CP or hybrid SDN/NFV in the future 5GRAN network architecture.

Machine learning (ML) and Deep Learning (DL) algorithms are under the umbrella of Artificial Intelligent (AI) which is addressed to enable intelligent network infrastructure [23–25]. ML and DL are popular for massive traffic identification, different QoS/QoE classification, massive devices behavior classification, especially for HetIoT devices, data profiling, RRH behaviors classification and statuses prediction, MEC server resource allocation, user Quality Class Identification (QCI), and so on. The classification of the user application, RRH, Evolved Packet Core (EPC) entities, and services of the MEC servers is mandatory to enable E2E NS. Without classification, different resource requirement applications and dedicated physical or virtualized computing resources are unable to establish NS methods. Due to the offered computing resource from MEC, the SDN controller will be possible for computing the learning model of ML/DL algorithms.

2.2 Challenging Issues

Although there are several contributions of MEC, SDN, NFV, ML/DL, and NS in the future 5G communication systems, there still exist many challenging issues that require effective handling methods. The insufficient control of radio resources, power, and energy issues occur when there is an increasing delay in the communication networks. Especially, there is the heterogeneity of VM and MEC servers in the ECN, which requires different power and resource management. The effective management policy of energy is important for profitable utilization. Even though multiple MEC servers are handling massive requests from different application users, there is insufficient real-time user application classification and lacking high accuracy recommendation towards the best MEC servers for specific services to the specific application users. The communication QoS/QoE is decreased due to the unspecific QCI applications, and poorly matching between request and serving entities. In Cloud Radio Access Network (CRAN), the base stations connect to the ECN. So the failure response increment of the ECN entities happens when the requests exceed the remaining capacity. Moreover, the service failure happens when the incoming request is concurrently attempted to connect to the inadequate capacity servers. A high accuracy failure prediction method based on AI system is required to provide an efficient recommendation system for matching the incoming traffic with the serving entities. Moreover, the communication in CRAN system is massive and the user requests are imbalanced.

VNI requires many VM for dedicated applications, while multiple VM offloads for computing and the excellent resource scheduling of each VM are crucial for power consumption [26]. The offloading computation from the physical to the VM, and from the VM to the physical computing load will take a duration that affects the communication latency. For lower power IoT network, the real-time application requires to handle the offloading method that reduces the loading time to overcome the communication delay, jitter, and also the communication power consumption of the IoT devices. The AI algorithms are proposed to cope with the offloading times and efficient resource allocation issues in massive VM. However, there are many crucial issues in applying ML/DL in the communication systems. The appearance of ML/DL for intelligent resource handling (see Fig. 3) suffers from the frequent change of the communication traffic, making it hard to create the training dataset and select the fit learning model to overfit the testing for boosting the model accuracy. However, the ML/DL model takes an amount of time for processing in the learning model. So, the real-time IoT network requires an excellent learning model to compute in a short time and also requires high computing power of MEC servers.

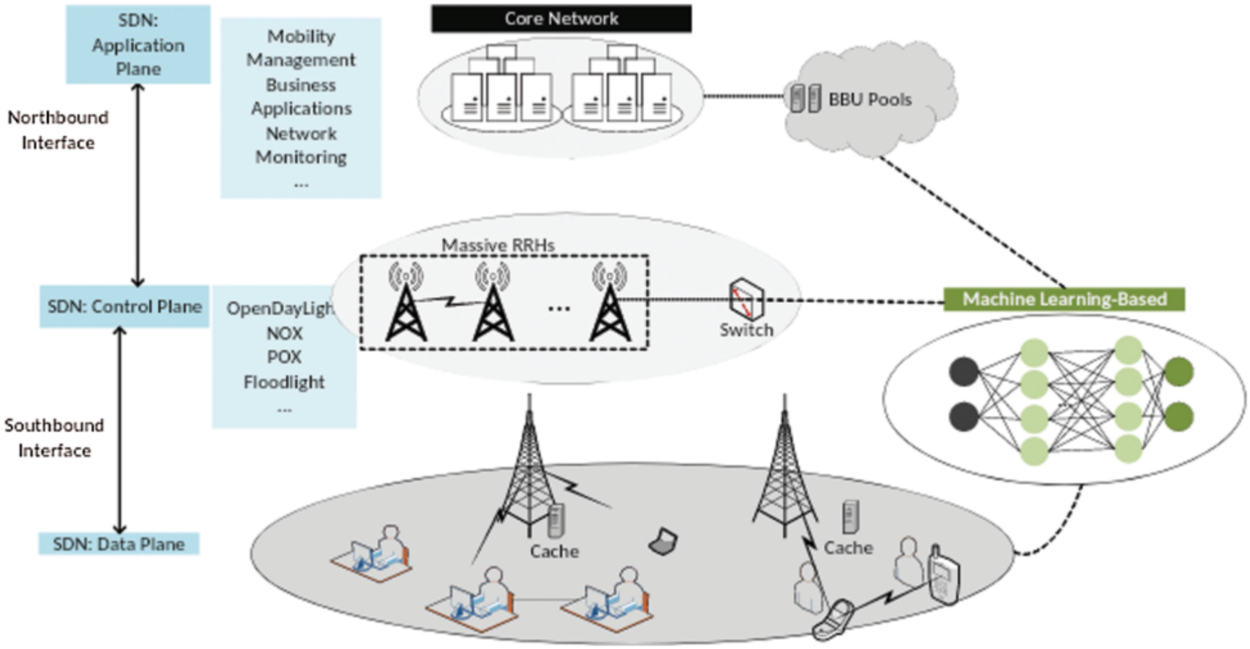


Figure 3: The intelligence radio network in 5GRAN architecture

Moreover, radio communication suffers from transmission drop as one of the open issues that are required to be solved. Each RRH has a limitation of a radio resource block for serving the massive IoT devices when they request to join the radio network at the same time. Whenever the RRH serves the massive IoT devices, the arising delays are incurred by the Transmission Time Interval (TTI) of the scheduling MAC to the communication channel. So, the transmission drops ratio increases based on the TTI. To enhance QoS/QoE at the radio network, an efficient method that can reduce TTI delays is required to improve communication reliability for real-time networks.

3 Proposed Scheme

In this section, the proposed lightweight ML-based real-time massive IoT traffic handling in 5G radio networks and the contribution of key technologies, including K-Nearest Neighbor (KNN), SDN, and the computed results caching is thoroughly discussed as follows.

As mentioned, the transmission drops in 5GRAN have become the main consideration point for real-time IoT traffic. This paper proposed an intelligent traffic handling based on lightweight

ML called KNN for classification and recommendation of the incoming traffic to meet the condition of multiple RRH gateways and balance the massive traffic to match with the availability of the obvious base station statues. The proposed scheme is executed by the converged computation of the aforementioned key technologies. MEC was proposed to cache the IoT traffic and RRH information. Also, MEC provides the computing power for the ML model which locates in the control plane infrastructure. The target of this paper is to focus on improving QoS/QoE of the real-time IoT communication in the edge networks, so a lightweight ML algorithm is suitable for the computing resources for operating in the ECN environments. ECN entities suffer from computing power because of the insufficient capacity of MEC servers. However, lightweight ML can be processed in a regular machine. The learning model is executed in a short period which is considerable for real-time applications. Supervised learning, KNN algorithm, was selected for classification and recommendation of the IoT traffic. KNN is widely used ML model for efficient recommendation systems. The SDN monitored and gathered the UP information, IoT traffic, and RRH statuses. While the CP is the brain of the proposed scheme, the learning model of ML took place in CP. Also, the output of KNN results was cached to the MEC servers and updating information was managed by the controller module. The scheme comprised of three stages, which includes the MEC part (store the information of IoT traffic, RRH status, and output results of KNN), ML models (classification and recommendation of the user traffic to the RRH), and SDN controller (monitoring, flow configuration, and system controlling).

The first stage is to balance the requests to fit with the RRH capacity or sink node statuses. The information of RRH has to be real-time monitoring, and the incoming traffic has to be categorized into different classes. Each class contains different numbers of traffic and recommends a specific serving gateway. Each gateway also receives the request traffic according to its status. To perform the caching of UP information and RRH statuses, MEC was proposed to be the caching servers that store the information of both UP and CP (output of ML). The computation of the KNN algorithm is also provided by the MEC server.

In the second stage, the precise training dataset is required for the learning model. The labeled dataset (training dataset) guides the IoT user traffic to match with the RRH gateways. The sufficient labeled datasets require to ensure that it contains enough information for the testing model. In case of insufficient information or too small volumes of the training dataset, the model evaluation will perform with low accuracy that the classification and recommendation will be unsatisfied. Whenever the ML model output poor evaluation scores, it provides unsatisfied QoS/QoE for the communication systems. In this paper, the KNN classified the IoT traffic and recommended to individual RRH gateways depending on three behaviors of RRH in terms of the delay (D) that occurred during the MAC-to-channel scheduling, delay D_{s1} that occurred at the S1-uplink interface and the transmission power (E) of the gNB. The KNN algorithm classifies and recommends the traffic to a specific gNB based on the *Dist* metric values. The entire operation flows of KNN (see in Fig. 4) are based on the training dataset, the gNB with higher transmission power E and lower delay D is recommended. The incoming traffic connects to the gNB with the lowest *Dist* metric as defined in Eq. (1) below. The gNB with a high *Dist* metric indicates that it is handling massive traffic, so D possibly arises. gNB with a higher D metric value requires to restrict the joining of IoT traffic. So, it is important to identify the gNB with a lower D metric and configures it to serve more incoming traffic than the other gNB.

In the final stage, the management and orchestration of the computing entities for caching incoming user traffic information, gNB statuses, classification, recommendation results of the ML algorithm, and the synchronous between UP and CP have to be performed by the SDN controller.

The output of the KNN guides the incoming traffic, SDN controller configured each of the IoT devices based on the recommendation of the KNN. The dynamic configuration between the IoT devices and gNB for radio communication is required. However, it is difficult for guiding the connectivity in the radio networks, because the IoT devices will automatically attempt to join the RRH with close distances.

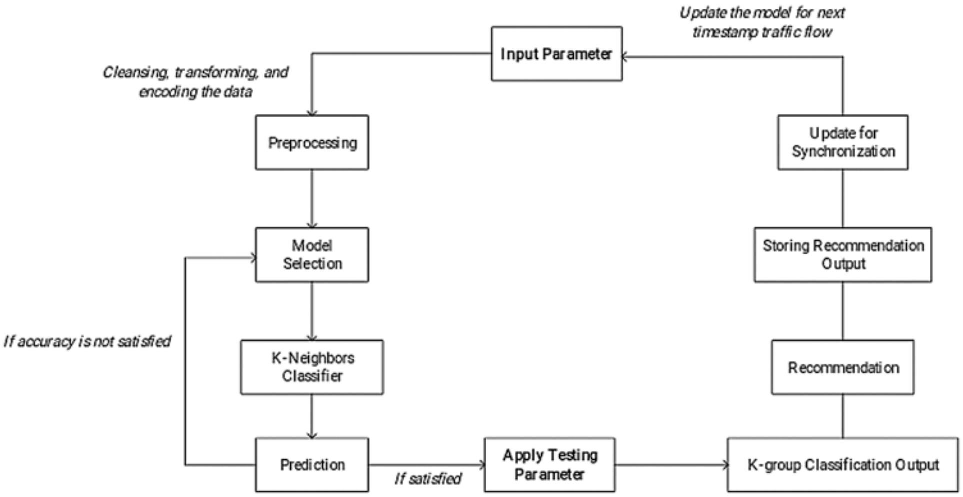

Input Parameter
Update the model for next timestamp traffic flow
Cleansing, transforming, and encoding the data
Preprocessing
Update for Synchronization
Model Selection
Storing Recommendation Output
If accuracy is not satisfied
K-Neighbors Classifier
Recommendation
Prediction
If satisfied
Apply Testing Parameter
K-group Classification Output

Figure 4: The classification and recommendation flow of the KNN ML

The *Dist* metric evaluation can be modeled as the expression (1) below.

$$Dist(D_i, D_{s1i}, E_i) = \sqrt{\sum_{i=0}^{n} \left(\frac{D_i + D_{s1i}}{E_i}\right)^2}, \quad E_i \neq 0; \quad i = 1, 2, 3, \ldots, n \tag{1}$$

where,

- *Dist* represents the status of the gNB with the relation of transmission power E, the delay D which occurred at the gNB physical interface and the D_{s1} is the delay that occurred at the S1-uplink interface. The gNB with the lowest number of *Dist* metric will be selected to serve for the incoming traffic. So, the gNB with higher transmission power E, lower delay D and D_{s1} will be received more amount of the request traffic.
- D_n represents the delay that occurred when the massive incoming traffic needs to be scheduled at the gNB interface. When delay at the gNB interface is increasing, the failure of the radio resource joining will occur and the increasing delay interval in the radio area will reduce the communication QoS/QoE. For real-time applications, it is mandatory to reduce the delay interval in the gNB for lessening transmission time and improving communication reliability.
- D_{s1} represents the delay that occurred at the S1-uplink interface. Therefore, the queuing of incoming traffic arises at S1-uplink interface, so, the delay D_{s1} will be increasing simultaneously.
- E_n refers to the transmission power in dB or signal strength. When there are many requests of IoT devices to a single gNB, the signal strength will be reduced. The lower power E_n

impacts the capability of transmission traffic. The gNB with higher metric values has a higher chance to be selected.

4 System Analysis

4.1 5G Communication Latency

The E2E latency of packet transmission in 5G communication systems can be written as T which is the addition of multiple delay occurrences, i.e.,

$$T = T_{Radio} + T_{Backhaul} + T_{Core} + T_{Transport} \tag{2}$$

where,

- T_{Radio} is the broadcasting radio delay which occurred at the interval communication of end sending devices and RRHs.
- $T_{Backhaul}$ is the packet transmission delay which occurred at the interval of gNB and the backhaul networking, particularly, the switching process at the Service Gateway (SGW) and Packet Data Gateway (PGW). The popular connectivity technique of gNB and the core network is a physical fiber-optic or microwave transmission link system.
- T_{Core} is the time taken within the core gateway connectivity establishment which is contributed by the control plane and data plane. For the control plane, there are several occurrence delays in various EPC entities including, the SDN controller, Home Subscriber Server (HSS), Mobile Management Entity (MME), and Policy and Charging Rule Function (PCRF).
- $T_{Transport}$ is the time taken within the remote network data transmission communications which is relied on four main metrics such as, distance, routing, link bandwidth, and switching protocol.
- The thorough broadcasting radio delay can be described in the following equation.

$$T_{Radio} = t_{FA} + t_{EU} + t_{tx} + t_{bsp} + t_{mpt} \tag{3}$$

where,

- t_{FA} is the time taken within the process of frame alignment, configuration, modulation, etc.
- t_{EU} is the synchronization delay between IoT devices and (R)AN gateways.
- t_{tx} is the packet transmission delay which relied on payload size, channel condition, and transport protocol.
- t_{bsp} is the time taken within eNB regions.
- t_{mpt} is the time taken which relied on the capacity of end devices and gNB terminal.
- The thorough $T_{Backhaul}$ can be described in the following equation.

$$T_{Backhaul} = t_Q + t_e + t_{tx} + t_s \tag{4}$$

where,

- t_Q is the time taken when the packet transmission waits within the queue.
- t_e is the time taken within the circuit performances of any network devices.
- t_s is the time taken within the switching operation.
- The thorough T_{Core}, the time taken within the core gateway connectivity establishment, can be described in the equation below.

$$T_{Core} = t_Q + t_e + t_{tx} + t_{epc} + t_{sr} \tag{5}$$

where,

- t_{epc} represents the delay of EPC entities' communication interfaces including, MME, HSS, and PCRF which is approximately within microseconds.
- t_{sr} represents the delay of the switching and routing operations.

The conditions of each RRH depend on the serving of each gateway and the network device from the heterogeneous network can be defined as the M/M/1 queue model as below.

$$\varrho = \frac{\lambda}{\mu} \tag{6}$$

where ϱ denotes the ratio of serving rate, λ denotes the incoming rate of user traffic and μ is represented the serving rate of the gateway or any base stations.

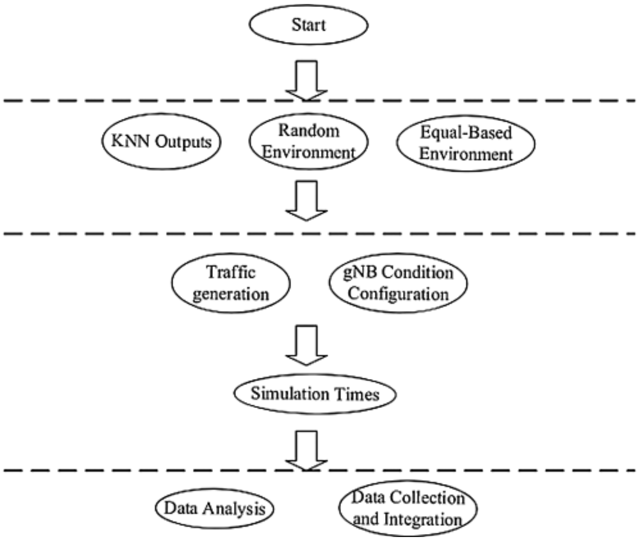


Figure 5: The simulation diagrams

4.2 Simulation Environment

The simulation system comprises of two different scenarios, including conventional (random and equal-cost based methods) and proposed approach based on KNN lightweight ML algorithm (see Fig. 5). There is 5,817 user traffic which was generated from the user devices at the same time. Four RRHs or gNB were used to conduct the experiment. There is 5,817 traffic in the dataset which was generated by Python programming. The dataset was generated based on the definition of 4 different RRH behaviors. The RRH behaviors were generated to reflect the real-world RRH statuses in massive IoT environments. RRH behaviors are based on transmission power E, scheduling delay between MAC and communication channel D, and delay occurrence at the S1-uplink interface D_{s1}. The training dataset was made by giving more weights to RRH with higher E, lower D, and D_{s1}. According to the trained dataset, the RRH with the lowest *Dist* metric value was considered as the optimal RRH (see Eq. (1)). Consequently, the ML model selected the optimal RRH based on the guided features of the trained dataset. The conventional approaches were simulated base on random access of incoming user traffic to the RRH and equal traffic handling of each RRH. For the random scheme, 5,817 was randomly accessed to 4 RRHs. In terms of the equal cost-based handling, the model separated user traffic equally for each RRH. While with the proposed scheme, each RRH was configured to serve the user traffic based on the output of KNN algorithm. The training and testing model of KNN was conducted by using the opened machine learning library, namely Sci-kit Learn. The real-time network simulation was conducted by using the discrete Network Simulation version 3 (NS3). The simulation time was set to 200 s, 40 user devices, and 3,332,250 total transmission traffic. And the Random Early Detection (RED) queue was integrated into the SGW/PGW gateway for buffering and QoS evaluation purposes.

5 Results and Discussion

In this section, results and discussion of the proposed and conventional schemes will be thoroughly presented as follows. The evaluation results were mainly based on the key important communication QoS parameters, such as average delay, average jitter, and average throughput in the radio network. The QoS evaluation at the S1-uplink interfaces presented the subjects of average communication delay and jitter. Moreover, the E2E communication reliability parameters in terms of packet (Pkt) delivery ratio, Pkt drop ratio, and Pkt drop count will be discussed in this section. The efficiency of the proposed scheme will be evaluated by comparing with the two conventional approaches, random and equal-cost methods.

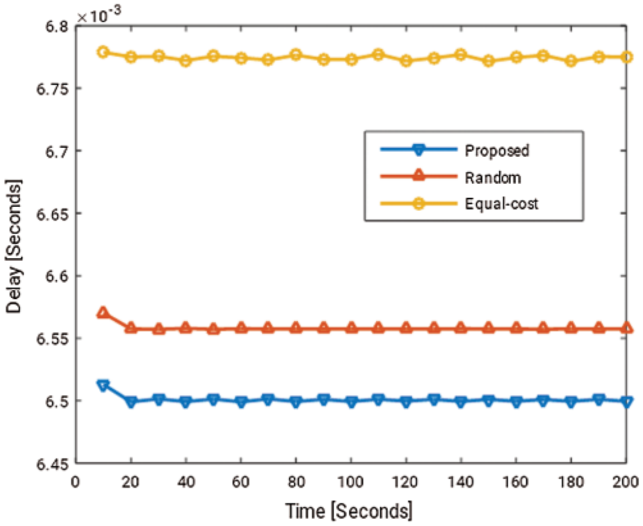


Figure 6: The average delays comparison between the proposed and conventional schemes

Fig. 6 depicts the average delay comparison of the proposed and conventional schemes in the radio networks. The graph shows that the proposed scheme noteworthy outperformed the conventional approaches due to the proposed scheme has lower average communication delays than random and equal-cost methods. The delay of the random method varied based on random situations, and if the random accessing of incoming traffic is more appropriate to the existed capacity of RRH, the communication delay will reduce. For a real-world scenario, communications can be randomized that the stability of communications will be inadequate. As shown in the graph, the random method has higher communication delays than equal-cost. And the equal-cost based communication has insufficient dynamic traffic handling, due to the RRH configuration for equal serving the user traffic. So, the delay can occur at the RRH with a lower capacity. The proposed scheme is used to improve the communication delay for real-time applications. By applying a lightweight ML for recommending the incoming traffic to the appropriate RRH, the ML offered high accuracy matching and assigned an appropriate amount of traffic to a specific RRH to be served. The reduction of TTI at gNB of the next generation RRH will be significantly mandatory for handling massive IoT traffic and improving the real-time application that required ULL. The capability of reducing communication latency improves not only the communication QoS/QoE but also the power consumption reduction of the IoT devices and lessens the device resource utilization in the communication systems.

The comparisons of average communication jitters between the proposed and conventional schemes in the radio networks are illustrated in Fig. 7. The graph indicates that the proposed scheme extraordinary outperforms the conventional approaches. And, the random and equal-cost schemes have higher communication jitter than the proposed scheme. The communication jitters represent the stability of the system when the interval of the metric jitter values (standard deviation of the delay) gets smaller and the stability of the communication gets higher. The ultra-high stability of communication is compulsory for some critical time-sensitive applications. According to the graph, the proposed scheme provides excellent system stability with minimum communication jitters. To improve communication QoS/QoE for user applications and to meet the perspective of 5G technologies, communication jitter minimization is an important key that has to be performed. So, the proposed scheme based on lightweight ML is suitable for handling the massive real-time IoT traffic in 5GRAN.

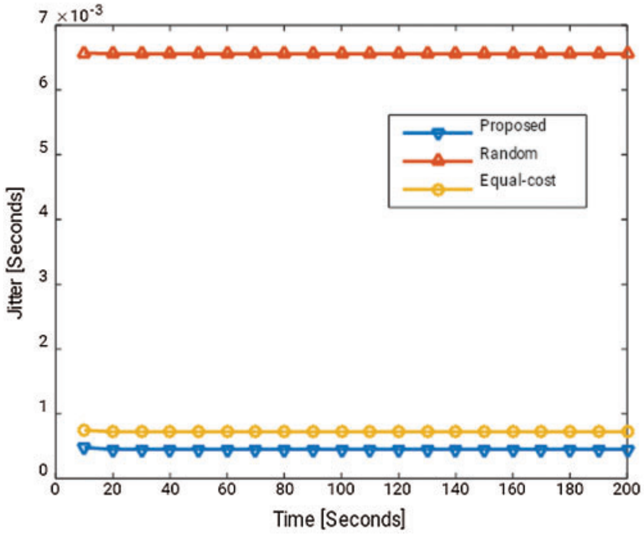


Figure 7: The comparison of average communication jitters between proposed and conventional approaches

Fig. 8 presents the comparison of averaged communication throughput between the proposed, random, and equal-cost methods. Based on the illustrated graph, the proposed scheme performed better than random and equal-cost. The proposed approach reached out up to 1,261.86465770492 Kbps of the average throughput, while the capable communication throughputs of random and equal-cost approaches were 1,250.8488806848 and 1,210.86474217841 Kbps, respectively. The greater capability of communication throughput will be a great opportunity for improving communication QoS and QoE. Higher communication throughput will reduce the E2E communication times of the user and the idle periods of the E2E network devices will increase. Regarding the mentioned benefits, enhanced communication throughput becomes a great occasion for boosting the battery life of lower power IoT devices. It is a great benefit for sensor IoT networks. The proposed scheme applied the KNN for inspecting the obvious RRH statuses. And, the curiosity of the RRH statues is convenient for the dynamic recommendation of the incoming user traffic to meet the appropriate RRH with sufficient resources for handling. The monitoring of RRH statues is mandatory for efficient control. The proposed scheme provided intelligent handling of massive IoT traffic in massive RRH networks and offered high accuracy of the recommendation by the lightweight ML algorithm. While the delay interval can be reduced by the proposed schemes, the capability of increasing the traffic delivery in the communication system is possible and the congestion that occurred during the flooding of massive traffic in radio gateways can be reduced.

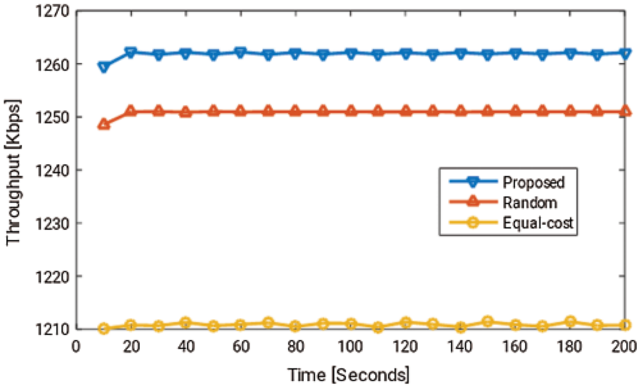


Figure 8: Comparison of the communication throughput between proposed and conventional (random and equal-cost) approaches

Based on the above-illustrated results, bad situations that occurred at the radio gateways are solved by the proposed scheme in terms of communication delay, jitter, and throughput. These significant QoS parameters have been improved by reducing the TTI at the physical RRH interface and S1-uplink interface. In the future, the 5GRAN network environment will handle the bigdata of user traffic generated by the mobile and HetIoT devices. So, heavy user traffic will become long queues at radio gateways. Due to the insufficient serving method of the default RRH, congestion will happen. In the case of inability to solve the congestion on times, the user traffic in the queue will be exponentially dropped. Some critical applications that required ultra-high communication reliability will be suffered from the congestion at radio gateways and the lessening of communication QoS/QoE will happen. The proposed scheme provided an efficient distributed control of massive IoT traffic in radio network environments that reduced bad network conditions. The total E2E packet drop count during the experiment of the proposed scheme outperformed the random and equal-cost methods (see Fig. 9). The total Pkt drop counts of the proposed scheme are only 560 of the 3,332,250 total transmission traffic, while the random and equal-cost Pkt drop counts are 770 and 940, respectively.

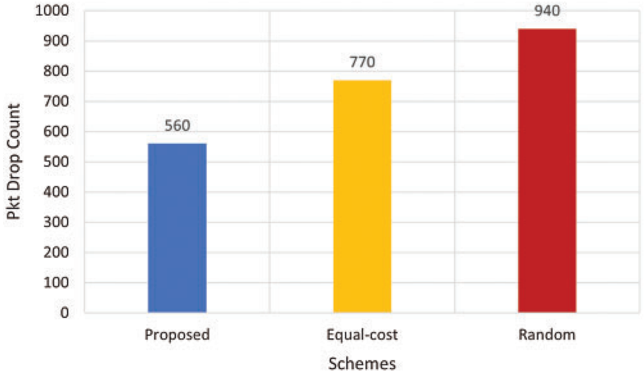


Figure 9: Total E2E Pkt drop count compared between the proposed, random, and equal-cost approaches

In the future 5G communication systems, the communication will be delivered mostly in the edge network area. While most OTT applications and services will be cached and located in edge network infrastructure. So, the congestion at backhaul and core gateways will be accordingly reduced. However, the significant issues in the backhaul and core network environment will migrate to the edge network area, especially the evolvement of massive traffic handling, heterogeneous MEC servers handling, massive heterogeneous RRH resource control, etc. To achieve the 5G QoS perspectives, the guarantee of QoS in 5GRAN has to consider. The proposed scheme mainly contributed to the radio gateways handling based on the KNN lightweight ML. The scheme considered on the RRH statues regarding TTI of physical gNB and the condition of the S1-uplink interfaces. Besides the enhanced delay, jitter, and communication throughput, the proposed scheme remarkably outperformed the conventional regarding E2E communication reliability as well.

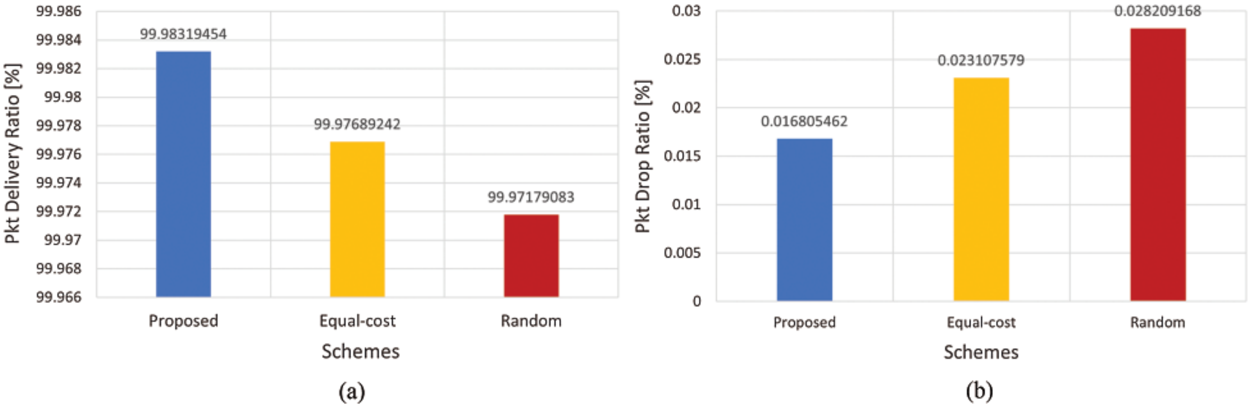
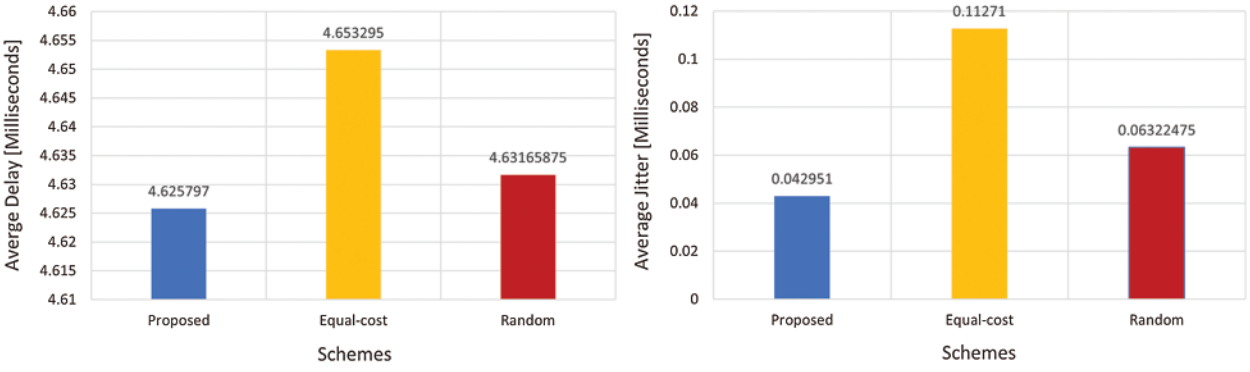


Figure 10: Presents the comparison of communication reliability between the proposed and conventional schemes. (a) The comparison of the Pkt delivery ratio between proposed, random, and equal-cost approaches. (b) The Pkt drop ratio comparison between the proposed, random, and equal-cost approaches

Fig. 10 represents the evaluation of communication reliability between proposed and conventional schemes. The transmission reliability of the proposed scheme reached 99.98319454% in 200 s of the simulation period (in the simulation environment, the reliable communication increased based on the cumulative simulation time), while the random and equal-cost schemes reached only 99.97689242% and 99.97179083%, respectively (see in Fig. 10a). Also, the proposed scheme had a lower Pkt drop ratio of 0.016805462%, while the random and equal-cost had a higher ratio of Pkt delivery drop ratio up to 0.023107579% and 0.028209168%, respectively (see Fig. 10b). Fig. 11 represents the comparison of average E2E delays and communication jitters of each user traffic between the proposed and conventional schemes, respectively. The average communication delay of the proposed scheme is better than random and equal-cost (see Fig. 11a). While, the average delay value of the proposed scheme is about 4.625797 ms only, and the random and equal-cost delays are 4.653295 and 4.63165875 ms, respectively. So, the proposed scheme outperformed the conventional approaches regarding communication jitters (see Fig. 11b). The average communication jitter of the proposed scheme is only 0.042951 ms, while the random and equal-cost jitters are up to 0.11271 and 0.06322475 ms, respectively.



(a)

(b)

Figure 11: Presents the comparison of E2E communication latency between proposed and conventional schemes. (a) The average E2E delay of each communication flow comparison between the proposed, random, and equal-cost approaches. (b) The average E2E communication jitters of each traffic flow comparison between the proposed, random, and equal-cost approaches

6 Conclusion

To guarantee E2E ultra-high communication reliability for massive real-time IoT and time-sensitive OTT applications in the 5GRAN environment, the efficient radio resources handling and dynamic gateways configuration of multiple RRH have to be considered. With the offered opportunities of convergence key technologies of ECN and high computing power of the MEC server, the lightweight machine learning algorithm was used for classification, prediction, and recommendation with a short period of computation. Whenever the RRH serves the massive IoT devices, the arising delay of TTI at the RRH and S1-uplink interface exponentially increased. This paper presented an intelligent traffic handling based on KNN lightweight ML to improve real-time communication QoS. The simulated results showed that the proposed scheme is remarkably outperformed the conventional approaches (random and equal cost) in terms of communication delay, jitter, throughput, and E2E communication reliability. Based on these great benefits, the proposed scheme is important and suitable for handling massive time-sensitive IoT traffic and improving the QoS/QoE. For further improvement on this research, we will focus on the deep packet inspection for the obvious status detections of the massive IoT packets towards effective dynamic resource provision.

Funding Statement: This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R1I1A3066543), and this work was supported by the Soonchunhyang University Research Fund.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat and H. Dai, “A survey on low latency towards 5G: RAN, core network and caching solutions,” *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3098–3130, 2018.

[2] F. Marzouk, J. P. Barraca and A. Radwan, “On energy efficient resource allocation in shared RANs: Survey and qualitative analysis,” *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1515–1538, 2020.

[3] J. Farooq and J. Soler, “Radio communication for communications-based train control (CBTC): A tutorial and survey,” *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1377–1402, 2017.

[4] M. Agiwal, A. Roy and N. Saxena, “Next generation 5G wireless networks: A comprehensive survey,” *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1617–1655, 2016.

[5] T. O. Olwal, K. Djouani and A. M. Kurien, “A survey of resource management toward 5G radio access networks,” *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1656–1686, 2016.

[6] M. Kamel, W. Hamouda and A. Youssef, “Ultra-dense networks: A survey,” *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2522–2545, 2016.

[7] I. A. Alimi, A. L. Teixeira and P. P. Monteiro, “Toward an efficient C-RAN optical fronthaul for the future networks: A tutorial on technologies, requirements, challenges, and solutions,” *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 708–769, 2018.

[8] S. Math, P. Tam, A. Lee and S. Kim, A NB-IoT data transmission scheme based on dynamic resource sharing of MEC for effective convergence computing. In: *Personal and Ubiquitous Computing*. Berlin, Germany: Springer Science + Business Media, 2020.

[9] Z. Li and Q. Zhu, “An offloading strategy for multi-user energy consumption optimization in multi-MEC scene,” *KSII Transactions on Internet and Information Systems*, vol. 14, no. 10, pp. 4025–4041, 2020.

[10] A. Rasheed, P. H. J. Chong, I. W. Ho, X. J. Li, W. Liu *et al.,* “An overview of mobile edge computing: Architecture, technology and direction,” *KSII Transactions on Internet and Information Systems*, vol. 13, no. 10, pp. 4849–4864, 2019.

[11] Y. Mao, C. You, J. Zhang, K. Huang and K. B. Letaief, “A survey on mobile edge computing: The communication perspective,” *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.

[12] J. Xie, F. R. Yu, T. Huang, R. Xie, J. Liu *et al.,* “A survey of machine learning techniques applied to software defined networking (SDN): Research issues and challenges,” *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 393–430, 2019.

[13] F. A. Lopes, M. Santos, R. Fidalgo and S. Fernandes, “Fernandes A software engineering perspective on SDN programmability,” *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1255–1272, 2016.

[14] C. Song, M. Zhang, Y. Zhan, D. Wang, L. Guan *et al.,* "Hierarchical edge cloud enabling network slicing for 5G optical fronthaul," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 11, no. 4, pp. 60–70, 2019.

[15] Z. Zaidi, V. Friderikos, Z. Yousaf, S. Fletcher, M. Dohler *et al.,* "Will SDN be part of 5G?," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3220–3258, 2018.

[16] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta *et al.,* "On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1657–1681, 2017.

[17] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang *et al.,* "A survey on mobile edge networks: Convergence of computing, caching and communications," *IEEE Access*, vol. 5, pp. 6757–6779, 2017.

[18] S. Kim and W. Na, "Safe data transmission architecture based on cloud for Internet of things," *Wireless Personal Communications*, vol. 86, no. 1, pp. 287–300, 2015.

[19] S. Kim, D. Y. Kim and J. H. Kim, "Traffic management in the mobile edge cloud to improve the quality of experience of mobile video," *Computer Communications*, vol. 118, pp. 40–49, 2018.

[20] C. Mouradian, D. Naboulsi, S. Yangui, R. H. Glitho, M. J. Morrow *et al.,* "A comprehensive survey on fog computing: State-of-the-art and research challenges," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 416–464, 2018.

[21] S. Math, L. Zhang, S. Kim and I. Ryoo, "An intelligent real-time traffic control based on mobile edge computing for individual private environment," *Security and Communication Networks*, vol. 2020, no. 3, pp. 1–11, 2020.

[22] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu *et al.,* "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201–1221, 2017.

[23] C. Zhang, P. Patras and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2224–2287, 2019.

[24] Z. M. Fadlullah, F. Tang, B. Mao, N. Kato, O. Akashi *et al.,* "State-of-the-art deep learning: Evolving machine intelligence toward tomorrow's intelligent network traffic control systems," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2432–2455, 2017.

[25] F. Pacheco, E. Exposito, M. Gineste, C. Baudoin and J. Aguilar, "Towards the deployment of machine learning solutions in network traffic classification: A systematic survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1988–2014, 2019.

[26] A. J. Gonzalez, G. Nencioni, A. Kamisiński, B. E. Helvik and P. E. Heegaard, "Dependability of the NFV orchestrator: State of the art and research challenges," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3307–3329, 2018.