8-16-2022 11:00 AM

# AI-Based Traffic Forecasting in 5G Network

Maryam Mohseni, *The University of Western Ontario*

Supervisor: Nikan, Soodeh, *The University of Western Ontario*
Joint Supervisor: Shami, Abdallah, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Engineering Science degree in Electrical and Computer Engineering

# Abstract

Forecasting of the telecommunication traffic is the foundation for enabling intelligent management features as cellular technologies evolve toward fifth-generation (5G) technology. Since a significant number of network slices are deployed over a 5G network, it is crucial to evaluate the resource requirements of each network slice and how they evolve over time. Mobile network carriers should investigate strategies for network optimization and resource allocation due to the steadily increasing mobile traffic. Network management and optimization strategies will be improved if mobile operators know the cellular traffic demand at a specific time and location beforehand. The most effective techniques nowadays devote computing resources in a dynamic manner based on mobile traffic prediction by machine learning techniques. However, the accuracy of the predictive models is critically important. In this work, we concentrate on forecasting the cellular traffic for the following 24 hours by employing temporal and spatiotemporal techniques, with the goal of improving the efficiency and accuracy of mobile traffic prediction. In fact, a set of real-world mobile traffic data is used to assess the efficacy of multiple neural network models in predicting cellular traffic in this study. The fully connected sequential network (FCSN), one-dimensional convolutional neural network (1D-CNN), single-shot learning LSTM (SS-LSTM), and autoregressive LSTM (AR-LSTM) are proposed in the temporal analysis. A 2-dimensional convolutional LSTM (2D-ConvLSTM) model is also proposed in the spatiotemporal framework to forecast cellular traffic over the next 24 hours. The 2D-ConvLSTM model, which can capture spatial relations via convolution operations and temporal dynamics through the LSTM network, is used after creating geographic grids. The results reveal that FCSN and 1D-CNN have comparable performance in univariate temporal analysis. However, 1D-CNN is a smaller network with less number of parameters. One of the other benefits of the proposed 1D-CNN is having less complexity and execution time for predicting traffic. Also, 2D-ConvLSTM outperforms temporal models. The 2D-ConvLSTM model can predict the next 24-hour traffic of internet, sms, and call with root mean square error (RMSE) values of 75.73, 26.60, and 15.02 and mean absolute error (MAE) values of 52.73, 14.42, and 8.98, respectively, which shows better performance compared to the state of the art methods due to capturing variables dependencies. It can be argued that this network has the capability to be utilized in network management and resource allocation in practical applications.



**Keywords:** 5G, cellular traffic forecasting, deep neural networks, big data, temporal analysis, spatiotemporal model

# Summary for Lay Audience

Forecasting telecommunication traffic is significantly important in providing intelligent management features as cellular technologies progress toward the fifth generation (5G) technology. It is critical to assess the resources need for each network slice and how they change over time since a substantial number of network slices are deployed over a 5G network. If mobile operators become aware of the cellular traffic demand at a certain time and location in advance, network management and optimization strategies will be more effective. The most efficient methods now dynamically allocate computer resources based on machine learning technology that forecasts mobile traffic. However, the accuracy of the forecasting model is vital. In this study, we focus on cellular traffic forecasting for the next day by using temporal and spatiotemporal approaches. In order to evaluate the performance of the various neural network models to forecast cellular traffic, a collection of real-world mobile traffic data is utilized in this study. It should be mentioned that the proposed spatiotemporal network has the capability to be used practically for resource allocation and network management.

## Acknowlegements

I would like to express my deepest gratitude to my Supervisors, Dr. Soodeh Nikan, and Dr. Abdallah Shami, who have led me to the field of AI-based traffic forecasting in 5G network, and have inspired me a lot in my thesis work. Their precious guidance and support during this thesis were so instrumental and helpful. I will forever be grateful for all the support given.
Also, I would like to extend my sincere thanks to the committee members, Dr. Ahmed Refaey Hussein, Dr. Shaimaa Ali, and Dr. Anwar Haque, for accepting to be on my committee.
Finally, I am thankful to my family and friends for the endless love and encouragement they have given me during my studies.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

**1D-CNN** one-dimensional convolutional neural network. ii, viii–x, 3, 41, 47, 55, 59, 60, 67, 68, 70–73, 79–82

**2D-ConvLSTM** 2-dimensional convolutional LSTM. ii, viii–x, 4, 41, 42, 62–65, 73–82

**3GPP** 3rd Generation Partnership Project. 15, 19, 26, 28, 33, 39

**5G** fifth-generation. ii, iii, viii, x, 1, 4–10, 12–23, 25–44, 80–82

**5GC** 5G Core. viii, 15, 31–36

**6G** sixth-generation. 21, 22

**AF** Application function. 7, 27

**AI** Artificial Intelligence. 5, 7, 12, 16, 17, 20, 21, 35

**AMF** access and mobility management Function. 7, 26–28, 33

**AN** access network. 26

**AR-LSTM** autoregressive LSTM. ii, viii–x, 4, 41, 47, 55, 62, 63, 68, 71, 73, 81

**AUSF** authentication server function. 7, 27, 28

**BS** base station. 18, 44, 46, 76, 77

**BSS** business support system. 32

**CAGR** compound annual growth rate. 1, 16

**CDR** call detail record. 46, 47, 73

**CITS** cooperative intelligent transport systems. 12, 13

**CN** core network. 26

**CNN** convolutional neural network. 44, 45

**CP** control plane. 33, 34

**CUPS** control and user plane separation. 25, 33, 34

**DN** data network. 7, 33

**eMBB** enhanced mobile broadband. 14, 21, 25, 28–30

**EPC** evolved packet core. viii, 30–36

**EPS** evolved packet system. 31–34

**FCSN** fully connected sequential network. ii, viii–x, 3, 41, 47, 55, 57, 58, 67–69, 71–73, 79, 81

**GRU** gated recurrent unit. 45

**IIoT** industrial internet of things. 14

**IoT** Internet of Things. 12–14, 29, 31

**ITS** intelligent transportation systems. viii, 12–14

**KPIs** key performance indicators. 21, 37

**L0** first level. 19

**LSTM** long short term memory. ii, 44, 59, 61, 62, 73

**M2M** machine-to-machine. 1, 16

**MAE** mean absolute error. ii, ix, x, 64, 66–68, 71–75, 77, 78, 80–82

**MEC** multi-access edge computing. viii, 9–12, 32, 38–40

**ML** Machine Learning. 16, 37, 41, 43

**MME** mobility management entity. 26, 27, 33, 35

**mMTC** massive machine type communication. 22, 25, 28, 29

**MNOs** Mobile network operators. 1, 15, 82

**MSE** mean squared error. 66–68, 74, 75

**NAS** non-radio signalling. 27

**NEF** network exposure function. 28

**NF** network function. viii, 23–28, 32

**NFV** network functions virtualization. 9, 10, 12, 15, 25, 39

**NG-RAN** next generation radio access network. 27, 32

**NR** new radio. 32–34

**NRF** Network Repository Function. 23, 28

**NSA** non-standalone. viii, 30, 31, 34

**NSSF** network slice selection function. 7, 28

**PCF** policy control function. 7, 27, 28

**PDN** packet data network. 33, 34

**POI** point of interest. 46, 76, 77

**QoS** Quality of Service. 12, 27, 37, 42, 80

**RAN** radio access network. 5, 7–10, 12, 30, 32–34, 38, 39

**RELU** rectified linear unit. 57, 59, 67, 68, 75

**RMSE** root mean square error. ii, 64, 66, 71, 72, 75, 77, 78, 81, 82

**RNN** recurrent neural network. 44, 59, 62

**RSUs** road side units. viii, 13

**SA** standalone. viii, 30–32

**SBA** Service-Based Architecture. 23–25, 30, 32

**SBI** service-based interfaces. 9, 23, 24

**SDN** software-defined networking. 9–11, 25

**SL** Supervised learning. 37

**SMF** session management function. 7, 27, 33–35

**SON** Self Organized Networks. 19, 43

**SS-LSTM** single-shot learning LSTM. ii, viii–x, 3, 41, 47, 55, 61, 67, 68, 70–73, 81

**STCNet** Spatial-Temporal Cross-domain neural Network. 46, 76, 78

**UDM** unified data management. 7, 27, 28

**UE** user equipment. 7, 26–30, 38, 39

**UP** user plane. 33, 34

**UPF** user plane function. 7, 27, 33, 34

**URLLC** Ultra-reliable low-latency communications. 14, 21, 25, 28, 29

**V2I** vehicle-to-infrastructure. 12, 13

**V2V** vehicle-to-vehicle. 12, 13

**V2X** vehicle-to-everything. 12, 13

**vEPC** virtual evolved packet core. 23

**VMs** virtual machines. 15

# Chapter 1

# Introduction

## 1.1 Motivation

One of the main factors driving the increase in mobile traffic globally is the constantly expanding number of wireless devices connecting to mobile networks throughout the world. Devices and connections are expanding more quickly (10% compound annual growth rate (CAGR)) than both the population (1.0% CAGR) and Internet users worldwide (6 percent CAGR). The average number of devices and connections per home and per person is rising, and this trend is driving that growth. The expansion of devices and connections is being significantly aided by an increasing variety of machine-to-machine (M2M) applications, such as video surveillance and healthcare monitoring. When it comes to devices and connections, M2M connections will have the highest growth, nearly doubling in size throughout the projected period (19 percent CAGR) to reach 14.7 billion connections by 2023. As shown in to Figure 1.1, smartphones will rise at the second-fastest rate, with a 7 percent CAGR (a 1.4-fold increase) [1].

Mobile network operators (MNOs) are expected to enhance and optimize system performance due to the growing volume of exchanged data and new requirements for higher peak data rates, enhanced reliability, and decreased latency facilitated by fifth-generation (5G) networks. Maintaining fast and reliable cellular connections is a key challenge for MNOs given the faster 5G networks and various connected devices. MNOs are required to utilize more

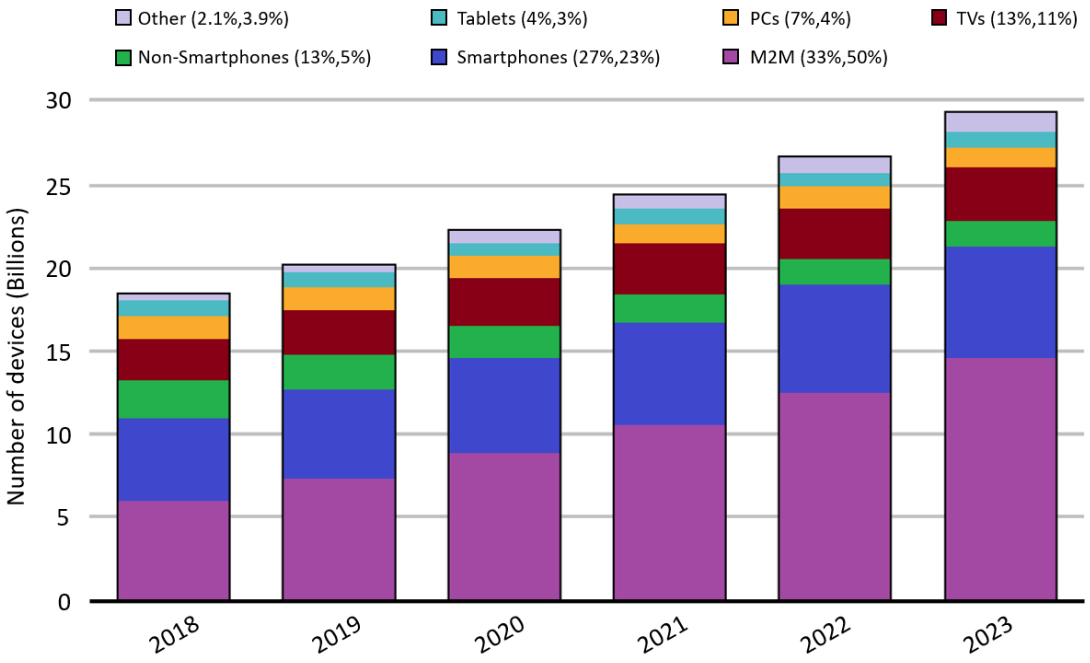


Figure 1.1: Global device and connection growth[1].

computing resources and expand the infrastructure. However, the risk of ineffective resource consumption is increased which might result in imposing extra costs on service providers. Network operators can expand their networks and allocate resources more efficiently and provide higher-quality services to their customers if they are aware of the demand growth in advance. The ability to predict mobile data traffic is crucial for resource management, reducing maintenance and operational expenses, and also meeting user needs.

Time, space, or a combination of two of them can be used to categorize data traffic prediction. While the spatial domain explores the data traffic based on the geographical loacation, the temporal domain investigates the trend and seasonality of data traffic. Also, the spatiotemporal domain captures both the temporal and spatial dependencies simultaneously. The major motivation of this study is to deploy machine learning techniques to extract the temporal and spatiotemporal dependencies of data traffic and forecast future cellular traffic. In this study, we used the telecommunication records of Milan from the Telecom Italia [3] dataset to predict cellular traffic. In fact, from temporal and spatiotemporal aspects, Milan cellular traffic is investigated using a variety of neural network models.

## 1.2 Problem Statement

Network management is a crucial step that relies on the current traffic demand and resource utilization in order to address the increment in cellular traffic. Accurately anticipating data traffic demand at a certain time and location with an appropriate strategy is required to satisfy the needs of the rapidly increasing data traffic.

Forecasting cellular traffic is a challenging task due to non-stationary, non-linearity and complex data dependencies and spatial and temporal correlation in the multivariate data. Diverse regions of the city, encounter a large dynamic range of data consumption due to various user involvement during work and leisure at different times of the day and on weekdays and weekends, leading to complicated temporal data patterns. Moreover, user movement and various population densities in different locations contribute to the development of spatial patterns. Predicting mobile traffic involves assessing the volume of cellular traffic for the next hours based on the preceding hours. Our goal in this study is to automatically forecast the 24-hour volume of telecommunications traffic using historical data from the preceding period of 24 hours. In the temporal domain, we aim to forecast the upcoming 24-hour volume of telecommunication, including sms, calls, internet, count, and frequency features. Furthermore, the spatiotemporal framework is intended to predict various sorts of cellular traffic, such as internet, sms, and call. The predictions that are a temporal sequence of network activity volume in various geographic areas are the main focus of the spatiotemporal framework.

## 1.3 Objectives and Contribution

The main objective of this thesis is to analyze the performance of machine learning in predicting the network traffic in high dimensional spatial-temporal cellular data. The goal is to investigate the effectiveness of various neural network models in traffic prediction from the univariate and multivariate perspectives. First, the univariate analysis will be conducted by applying the temporal framework including the fully connected sequential Network (FCSN), one-dimensional convolutional neural network (1D-CNN), single-shot learning LSTM (SS-

LSTM), and autoregressive LSTM (AR-LSTM) to capture the temporal dependencies. In the second section, the multivariate spatiotemporal analysis will be conducted using 2-dimensional convolutional LSTM (2D-ConvLSTM) to forecast the traffic in Telecom Italia [3] data in city of Milan. The objective in the multivariate spatiotemporal analysis is to automatically incorporate the inter dependencies among different variables, spatial and temporal information into our predictive modeling. Our objective is to dvelop a model that has the potential to aid in network management and resource allocation in 5G networks.

The main contributions of this work can be summarized as follow.

- This study focuses on predicting the cellular traffic of the next 24 hours by utilizing both temporal and spatiotemporal analysis.
- The proposed temporal and spatiotemporal models can be used to facilitate resource allocation and network optimization in the 5G network by accurately predicting the cellular traffic.
- We introduce a new feature named "count," which is utilized as model input. Since this metric shows the number of records in a certain period of time for a specific grid id, using count as model input aids in the prediction of various kinds of cellular traffic.
- The multivariate spatiotemporal analysis takes the variables' correlation and spatial and temporal dependencies into account in the predictive modeling.
- The proposed models have relatively low complexity, small number of parameters and short execution time for forecasting the traffic of the internet, call, and sms.

## 1.4 Thesis Structure

The remaining of this thesis is organized as follows:

- **Chapter 2** presents the literature review including the vision and motivation of 5G networks, network slicing, 4G to 5G migration, significant technologies, emerging use cases of 5G, and challenges and opportunities of 5G.

- **Chapter 3** provides background information on the 5G core. This chapter covers the concept of service-based architecture, main network functions, network slices, 5G core deployment approaches, the 4G to 5G core migration, role of artificial intelligence (AI), edge computing, and next-generation radio access network (RAN).
- **Chapter 4** discusses the related works on the study of mobile network traffic and also explains the utilized dataset. Also, in the methodology section of this chapter, preprocessing, temporal and spatiotemporal predictive models, and results and evaluation are elaborated.
- **Chapter 5** as the final chapter is devoted to discussing the advantages and drawbacks of the proposed models as well as drawing the conclusion, and outlining the future works.

# Chapter 2

# Literature Review

## 2.1 Vision and Motivation

5G is a unified, more capable air interface. It has been designed with an extended capacity to enable next-generation user experiences, empower new deployment models, and deliver new services. 5G wireless technology is meant to deliver higher data speeds, ultra-low latency, more reliability, massive network capacity, increased availability, and a more uniform user experience to more users. Higher performance and improved efficiency empower new user experiences and connect new industries. The industry is envisaging to see how networks might be utilized to address future intense capacity and performance needs as the demand for improved mobile broadband experiences continues to increase. Enabling a host of diverse platforms to function together as a unified entity, mostly software-controlled and adaptable to any consumption pattern, is the actual challenge. In this context, 5G is expected to meet industrial and social demands. It focuses on improving capacity by combining existing methodologies with advancements in radio technology or if necessary, a transformation in system design concepts. Also, 5G has delivered fast and pervasive internet coverage since 2020. Standardized and more uniform solutions would allow for substantially bigger volumes and hence higher integration densities. In addition, the proposed solutions will lower energy usage, and reduce expenses [4]. One of the key enablers for 5G networks include the deployment of artificial intelligence

(AI) .

## 2.2 Architecture

Control plane operations, user plane functions, and subscriber data management are all supported by a 5G network. In fact, mobile devices have been acquainted with the logical depiction of mobile network architectures since the beginning of the global system for mobile communications (GSM) and subsequently general packet radio service (GPRS). These schematics are composed of functional blocks that are connected to each other. The reference point representation of 5G architecture is shown in Figure 2.1. In this architecture, the control plane and user plane functions are separated, with the control plane being further divided into subscriber management and control plane functions. The subscriber management functions include the unified data management (UDM) which is the development of the home subscriber service (HSS), and the authentication server function (AUSF).

As for the control plane, a session management function (SMF), access and mobility management Function (AMF), a network slice selection function (NSSF), a policy control function (PCF), and an application function (AF) make up the control plane function. In terms of the user plane, there is user equipment (UE) which might be a smartphone.The UE links to the user plane function (UPF) and then to a data network (DN) including a business intranet and a public internet, through the RAN. Also, the connection among the UE and RAN is established through the air interface [5].

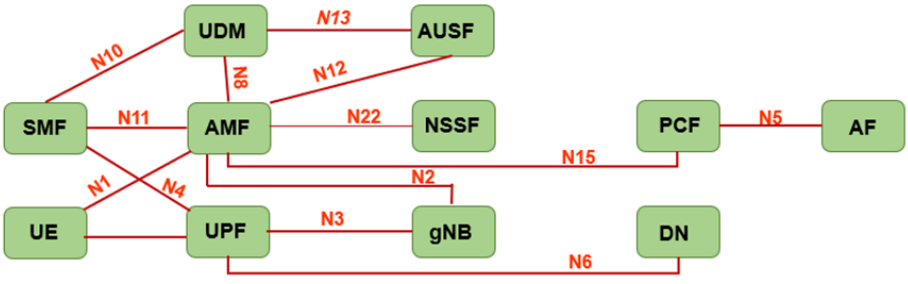


Figure 2.1: 5G Network Architecture.

## 2.3 Network Slicing in 5G

The splitting of a physical network into numerous virtual networks, each of which may be modified and optimized for a certain kind of application , is referred to as network slicing. The shared physical network resources may be dynamically and effectively allocated to logical network slices depending on changing user needs by leveraging cloud computing and virtualization technologies. Network slicing in 5G is made up of a set of network functions that have been integrated for a particular use case [6].

### 2.3.1 Challenges in network slicing

As for sharing radio resources among RAN slices, an apt radio scheduling method is required for avoiding certain issues. Also, mobility management in network slicing such as smooth handover creates some challenges. For real-time services, quick mobility handover is critical, and this has a direct impact on service quality. Another challenge relates to the mobility needs of network slices since the mobility management requirements vary. For example, the mobility requirements of the mobile broadband slice differ from those of the autonomous driving slice [7].

### 2.3.2 Deploying network slicing in 5G

In terms of slicing deployment in 5G, distinct services with diverse requirements including low latency and high motility should be combined into one shared physical infrastructure in 5G network. Also, 5G network supplies specialized network slicing for each of the services. A network slice is a virtual network that runs on top of a physical network, giving the tenant the impression that it is running on its own physical network. To satisfy a variety of service requirements, network slicing must be implemented in an end-to-end fashion. Also, protocols and network architecture may be distinct for every slice.

The process of slicing a 5G network comprises slicing the 5G core network, 5G RAN, and even user equipment. To construct core network slices for specific service demands, software

defined networking (SDN) and network functions virtualization (NFV) may be used to dynamically manage virtual network resources such as bandwidth and processing capacity. To apply the 5G RAN slicing, the logical abstraction to physical radio resources and physical hardware such as a base station should be implemented. It is possible to devote RAN and core network slices to a single class of service users or they can also be shared across many classes [7].

## 2.4 4G to 5G Migration

For vast majority of operators, the migration from 4G to 5G is a huge step since it affects numerous areas, and each of these areas must be planned and converted to take full advantage of 5G technology. As the 5G architecture [8] varies from the current 4G architecture, it is necessary to consider the following facets:

### 2.4.1 Preparedness in infrastructure and data center

5G requires a distributed data centre strategy, with some central data centres containing signalling and multiple core network functions, as well as numerous regional data centres holding multi-access edge computing (MEC) nodes. In addition, there might potentially be several other data centres hosting the dispersed RAN nodes [9].

### 2.4.2 Preparedness in radio network

Virtualization and separation of the 5G radio are required in this stage. The scattered and central components that may be disseminated over far edge data centers and software integrated with radio interface units (RIU) are included in the radio network nodes [9].

### 2.4.3 Preparedness in core network

The core of the 5G network differs significantly from prior generations. Furthermore, the messaging infrastructure must be modified for supporting cloud-native services as the major network elements are based on cloud-native and they leverage service-based interfaces (SBI),

for example the REST [9].

### 2.4.4 Preparedness in automation

Given the virtualization of all network functions in 5G, including RAN, and the sophisticated distributed data centers strategy, automation along with surveillance, administration, upgrading, and evaluation is essential for the deployment of multiple network functions [9].

## 2.5 Significant technologies in 5G

Novel concepts and approaches are being developed to create a network platform that is adaptable, agile, scalable,and programmable. Among these strategies, NFV, SDN, and MEC are the three main techniques used in 5G networks which are elaborated below.

### 2.5.1 NFV

NFV will enable network slicing in 5G, a virtual network architectural feature that allows many virtual networks to be constructed on top of a common physical infrastructure. For instance, the virtualized firewalls and load-balancers on commodity hardware are referred to as NFV[10].

In comparison with the traditional approach, NFV introduces the following three modifications in how network services are supplied [11].

- Decoupling software from the hardware platform: In NFV, the hardware and software entities are not coupled, and their operations can run individually in parallel to one another.
- Implementing network functionalities with more flexibility: Software and hardware may execute distinct functions at different times as they are separated. This allows operators to implement new innovative applications while still utilizing the same hardware platform.
- Providing services and network operation dynamically: By scaling the performance of NFV dynamically, network operators may launch customized services based on client requirements.

### 2.5.2 SDN

To overcome hardware restrictions, SDN has evolved as a new intelligent design for network architecture. SDN's major goal is to decouple the control plane from the switches and enable external data control via a logical software component known as controller. Simple abstractions are leveraged by SDN to elaborate the elements and their functions besides the protocols required for managing the forwarding plane and Mobile IP from a remote controller via a secure connection. Therefore, the difficulty to mutually access various sections of diverse networks would be resolved. For the majority of switches and associated flow tables, this abstraction is employed instead of the conventional methods such as forwarding tables. As a consequence, the controller examines network packets, resolves faults, and publishes policy based on monitoring outcomes.

Furthermore, since hardware components are costly, Internet service providers cannot afford to incur large upgrades, adaptation, or construction expenditures to meet the constantly expanding requirements of consumers. As a result, another benefit of utilising SDN is that it makes it simpler to create and deploy new products and services than traditional hardware-dependent standards.

Finally, the main purpose of SDN is to construct a network that does not require any administrator intervention in terms of design or changes. Hence, the network's management may be totally automated and will be supervised more effectively through the controller plane by dictating the required policies to the routers and switches while maintaining complete network monitoring. [12].

### 2.5.3 MEC

MEC provides cloud-based applications, IT, and network services at the mobile network's edge [13]. MEC also processes data near the point of production and consumption. This allows the network to provide the ultra-low latency essential for various applications while still supporting customer experiences in high-traffic areas. In addition, the cost of data transfer can greatly be minimized by processing locally in MEC applications [14].

MEC achieves numerous important network benefits including improved QoS to end-users in the context of video streaming allowed by 5G network slicing as well as optimization of mobile resources at the network edge [15]. Furthermore, the MEC application server, which operates on top of the MEC NFVI infrastructure and offers services to end-users as separate MEC Applications (MEC Apps), is the key part of MEC in an architectural examination of MEC and NFV. The MEC platform, where MEC services are maintained, shares communication interfaces with MEC Applications. MEC service nodes can execute locally or remotely in the cloud, depending on the installed data center. Interfaces to the traffic offload function (TOF), which are embedded in the data plane and prioritize traffic via visible packet monitoring, are included in both MEC Apps and MEC services. MECs' integration into the RAN is made easier as a result, and they play a crucial role as a general surveillance element [16].

## 2.6 Emerging use cases in 5G

### 2.6.1 Smart cities

Smart cities refer to an approach in which information, communication, and technology (ICT) are combined with a city's conventional structure and then organized and managed using digital technology. The sensors and actuators integrated into smart devices that detect the environment for effective decision-making are at the heart of the smart city. The microcontrollers in these devices have been designed to make decisions spontaneously depending on the data collected by sensors. This entails combining various information and communication technologies such as the internet of things (IoT), wireless sensor networks (WSN), AI, and protocols [17].

### 2.6.2 Intelligent transportation systems

5G aims to connect individual cars through the development of cooperative intelligent transport systems (CITS). The main technologies used for intelligent transportation systems (ITS) include vehicle-to-infrastructure (V2I), vehicle-to-vehicle (V2V), and vehicle-to-everything (V2X) [18]. In the V2I system, the traffic data created by the vehicle will be initially col-

lected. Then, warnings regarding environmental concerns or safety from the infrastructure will be transmitted to the automobile and will notify the driver [19].V2V communication allows cars to convey information such as their location and speed wirelessly, which can reduce traffic jams and minimize collisions [20]. Finally, V2X is a communication system that allows cars to interact with each other and with the infrastructure surrounding them, such as traffic lights. Thus, the key components of V2X are V2V and V2I [21].

To consider a large ITS, multiple road side units (RSUs) are essential, each of which handles many IoT devices and edges and connects with other RSUs for accessing data and performing analysis. Every RSU tends to cover a particular physical zone, and the devices in that region are usually associated with it. A more dynamic relation, on the other hand, is feasible. Since vehicles often employ short-range wireless communications, they are more likely to connect to various RSUs as they travel. The vehicular clients and roadside units in the ITS model which deploy a hierarchical edge computing approach with a three-layer design are depicted in Figure 2.2. Edge and IoT devices, such as sensors and cameras, comprise the bottom layer, which constantly feeds data to the monitored region. Then, a collection of RSUs in the middle layer streams data from the devices in its proximity. These RSUs store data for a limited time and analyze the real-time data to acquire intelligence. Ultimately, an ITS server (traffic control center) that enables long-lasting data storage and sophisticated data investigation is embedded in the top layer[2].

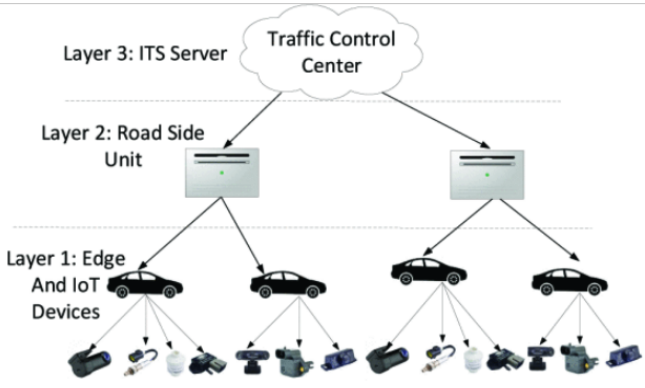


Figure 2.2: Vehicular clients and RSUs in ITS model deploying Edge Computing [2].

Consequently, cities with 5G-enabled CITS will be smarter, and automated transit systems

will be more effective and safer compared to the current networks. By providing a genuinely smart transportation system with high-speed Internet access in public transportation, 5G can cope with main transportation concerns in megacities, such as pollution, traffic jams, and accidents. Moreover, the real-time data from connected devices and vehicles such as cars can be gathered and assessed by a linked traffic system which aids in enhanced navigation, decision management, and resource optimization. In fact, the various slice types in 5G including massive Internet of Things (mIoT) and enhanced mobile broadband (eMBB) can support and improve numerous elements of ITS operation [22].

### 2.6.3 Internet of things (IoT) and industrial internet of things (IIoT)

IoT is a technological concept describing the pervasive Internet connection that transforms regular objects into linked gadgets. However, in most cases, it is best to think of IoT as consumer IoT [23] [24]. In the consumer IoT, "things" are smart consumer electronic gadgets that are connected. In fact, they can increase human awareness of their surroundings which leads to saving money and time. Also, all industrial assets, such as machines and control mechanisms are linked to business processes and information in the IIoT. The IIoT is the term given to the application of IoT in the industries [25]. A huge number of smart industrial machinery and sensors interact with one another to establish a network of smart IoT-based gadgets with processing capabilities, storage, and communication management [26], which may eventually be utilized to execute complex computations cooperatively.

The maximum potential of IoT, including improvement in efficiency and connectivity, and data speeds for industrial applications, can only become feasible if it is supported by a flexible communication network able to support a variety of requirements, ranging from ultra-reliable low-latency communications (URLLC) to massive connectivity [27]. Technological breakthroughs in 5G telecommunications have emerged as the heart of IIoT applications [28], delivering quicker data transfer, higher bandwidth, and better spectral efficiency, all backed by micro operators and localized private networks [29].

## 2.7 Challenges in 5G

### 2.7.1 Network softwarization

Network softwarization refers to the process of designing, implementing, operating, and maintaining network components using software based network functions instead of traditional hardware supported ones.

### 2.7.2 Delays in network softwarization

The 5G Core (5GC) Networks commonly utilize cloud-native software packages. Also, 5GC entities should be deployed in distinct virtual machines (VMs). Practically, 5GC is used to handle calls in a hierarchical manner at multiple local and national data centres. Also, in response to changes in incoming traffic, the orchestrator initiates automatic service in a virtual environment. Moreover, the VMs can be started or stopped by leveraging a maintenance management architecture. Therefore, various state of network can be created by scaling VMs which results in extra delays for generated traffic. Furthermore, the execution of functions on hypervisors that separate virtual resources from a commercial-off-the-shelf (COTS) causes systematic time delays. However, the overall required time for 5G services that has been influenced by this delay is minimal and it can be handled in seconds [30].

### 2.7.3 Complexity

The advance from 4G to 5G also makes the protocol more complicated. This complexity is due to the existence of many different sorts of commands (messages) transmitting among MNOs in 5G compared to 4G. This is derived from the 3rd Generation Partnership Project (3GPP) Rel-15, and the quantity of command and information components delivered both intra and inter of the MNOs for Rel-16 is much more significant than the previous releases [31].

Also, the use of data-based networking to simulate NFV-enabled networks necessitates massive quantities of network-generated data. As a result, the size of network-generated data grows which leads to network complexity [32].

The next factor which plays a vital role in causing complexity is the increase in the number of users and devices. Devices and connections are rising at a quicker rate denoted by a 10 percent CAGR compared to the population and Internet users combined (1.0 and 6 percent CAGR).

Every year, a variety of new gadgets with enhanced functionalities and intelligence are released and endorsed in the marketplace. An enormous range of M2M applications, such as healthcare control and video inspection are fuelling growth in the number of devices and connections. M2M connections will possess the quickest growth in the device and connection sector, rising approximately 2.4-fold during the prediction time frame (19% CAGR) to 14.7 billion associations by 2023. Whereas M2M traffic has primarily been less than the traffic from the end-user devices such as phones and PCs, it is estimated that the volume of traffic is expanding faster compared to the number of connections due to the higher implementation of video applications on M2M connections and increase in the applications like those of telehealth and smart vehicle navigation systems, which need more bandwidth and lower latency [1].

Moreover, the next-generation use cases which are developing would lead to further complexity. For instance, the huge volume of data produced by billions of locally linked devices in smart cities needs sophisticated data management and processing approaches, as well as deeper and broader insights. Also, technical roadblocks in the shift from the traditional to smart systems exacerbate the complexity. Scalability, backward compatibility, diversity of devices and data, interoperability, and multiple data standards, all bring problems and challenges that must be tackled [17].

### 2.7.4 Tackle the 5G complexity with machine intelligence

5G networks are intrinsically complex and will be implemented at a scale that will eventually outpace human operators' capacity to stay on top of network and service management in order to meet service level agreements (SLAs) and maintain the end user's quality of service.

The deployment of 5G will need a shift away from operator-centric network management. ML and AI will be considered necessary to supplement the capabilities of operations staff by

producing real-time operational intelligence that directs network automation and orchestration functions with the least manual interference. The functioning of cloud-native infrastructure is also dependent on machine intelligence. Continuous input on the system state is required for continuous integration/continuous delivery (CI/CD), which is obtained from analytics, machine learning, and AI, and this operational intelligence is utilized to drive automation and orchestration activities in the cloud-native architecture. Leveraging machine intelligence, will have a direct correlation in the reduction of 5G complexity [33].

### 2.7.5 Security

To minimize hacking threats, 5G cybersecurity requires major upgrades. Some of the security concerns stem from the network itself, whereas others are related to the equipment that connect to 5G. However, these issues endanger customers, and corporations. One of the key issues regarding security in 5G is decentralized security. The traffic routing points in 5G's dynamic software-based systems are considerably greater. All of these must be monitored to enable 5G to be totally secure. Considering this may be difficult, any unprotected locations may jeopardize other sections of the network.

Another concern worth mentioning is related to surveillance of security due to higher bandwidth. The advantages of a larger 5G network might be detrimental to cybersecurity. Because of the increased speed and volume, security teams will be forced to devise new techniques for countering attacks [34].

### 2.7.6 Mitigation for security issues

To address the specific security issues of 5G, network providers should pay more attention on software security. They will need to cooperate with cybersecurity companies to provide encryption solutions, network monitoring, and other services. Furthermore, it is essential to highlight the importance of cybersecurity by educating customers. So, consumers must be informed of the importance of keeping all internet-connected devices up to date with software upgrades. Along with the initial implementation of 5G, attempts to improve security are un-

derway. However, because we require actual information to fine-tune the protections, work is still ongoing long after 5G has been launched [34].

### 2.7.7 Big data in 5G

Numerous connected devices in 5G network, and multiple use cases and services generate big data which leads to more traffic. With growing number of devices, 5G faces challenges as it floods the network with a huge amount of data. Besides being large, this massive data is diverse as well since it is gathered from distinct sources [35].

### 2.7.8 Managing big data in 5G

Calls are established by sending forward and reverse call setup signals between the local and remote users. The voice or data call between the users will follow identical pathways after the call is formed; however, the remote and local base station (BS) will vary dynamically based on the motion of the two users. When a mobile phone passes from one BS coverage area to another, control is transferred from the prior to the latter using a technique known as handoff or handover, which leverages the relative intensities of the signals received at the two nodes to determine the control transfer decision. BSs, also known as access nodes (ANs), offer users access to the network's infrastructure. Since each cellular region is covered by just a single AN and can handle only a few carrier frequencies, this old design does not satisfy the 5G demand that a massive number of devices with high-bandwidth needs to remain connected in any area for a long period of time. To overcome this restriction, a radical transformation from AN-centric to user-centric design is required. This architecture is founded on the concept that providing every user/device with a personalized BS to achieve the objective of always-on connectivity without sacrificing quality or user needs [36]. The challenges of 5G are summarized in Table 2.1.

Table 2.1: 5G Challenges.

| Challenges in 5G | Definition |
|---|---|
| **Delays in Network Softwarization** | • Various state of network causes delays.<br>• The execution of functions on hypervisors results in time delays. |
| **Complexity** | • Various sorts of commands in 5G leads to complication.<br>• The size of network-generated data grows and increases the complexity.<br>• Increase in the number of users and devices causes complexity.<br>• Next-generation use cases add to complexity. |
| **Security** | • Security concerns stem from network itself and the equipment.<br>• Decentralized security which means the traffic routing points in 5G are greater. |
| **Big data** | • Growing number of connected devices floods the network with huge amount of data.<br>• The diverse massive data are gathered from different sources. |

## 2.8 Opportunities facilitated by 5G

### 2.8.1 Self-healing

The ability to deliver a seamless continuous service became a need after the emergence of the ubiquitous network utilising 5G technology. Given the massive size of the 5G network, diagnosing and maintaining the network is a challenge. As a result, an automated system capable of performing network diagnosis and predicting self-healing is essential.

The Self Organized Networks (SON) architecture can manage the automation of the fault tolerance diagnosis and the rectifying procedure [37]. The self-healing process comprises a variety of actions and functions that are necessary for the smooth recovery from various defects. Figure 2.3 depicts a flow diagram for a first level (L0) self-healing process model that includes five major sub-processes described as: monitoring and detection, diagnosis of fault, and system compensation and recovery [38]. The cycle of the self-healing process is represented in the process model. Every process in the model is designed for conducting one or several of self-healing functions specified in the 3GPP technical standard for self-healing principles and standards. L0 is a self-healing general process model that begins with evaluating the system and terminates with recovery in the event of a fault [39].

The research conducted by Peter Szilagyi and Szabolcs Novaczki presents an automatic recognition and detection framework in an experiment to identify the root cause of the failure or to offer the appropriate action to heal from the failure without being aware of the underlying

process[40]. To manage their impact on self-healing, the framework focuses primarily on the detection and diagnosis functions. In terms of getting the diagnosis, learning the effects of faults on various performance metrics and recognizing deviance from typical behaviour are used. To improve the performance of the diagnostic function, a decision support system (DSS) must be built to assist in the root cause analysis for each failure that occurs in the network [38].

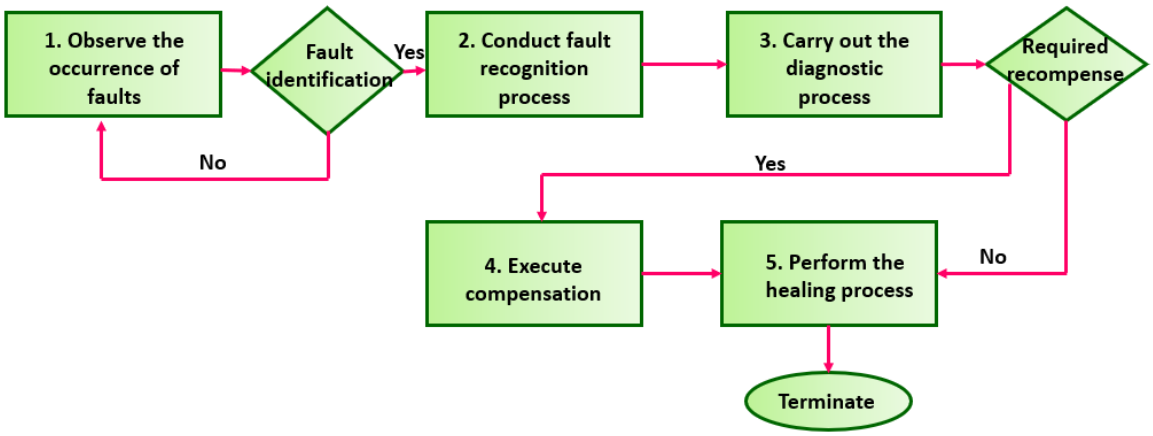


Figure 2.3: Self-healing model.

### 2.8.2 Self-configuration and self-optimization

Based on the information supplied by terminals as well as its own observation and cognition, the 5G network may execute intelligent configuration and optimization. For instance, the 5G network can discover new neighbour cells that have yet to be added to the neighbour list and perform automated neighbour cell addition. Also, the 5G network employs AI technology to create a model that reflects the association between the reason for radio link and handover failure and the radio link failure information given by the terminal and the handover record transferred between the base stations. 5G networks may use the above model to dynamically change parameters to prevent ping-pong handover, too early handover, and too late handover. Furthermore, the 5G network can recognize effective migration of users between base stations and minimize cell congestion by leveraging perception information transmitted among base stations such as resource usage and the number of active and connected users [41].

### 2.8.3 Network management

Maintenance and administration of mobile networks are still done manually or semi-automatically nowadays. Every mobile operator must have an operational group with a significant number of network specialists to diagnose faults, adjust software, and repair hardware during the troubleshooting process [42].

5G system must, on the other hand, become considerably more sophisticated and diversified than existing systems to fulfill various and radical KPIs of improved mobile broadband and Internet of Things. This places a lot of strain on today's network administration, which is already expensive, fragile, and time-consuming [6].

To reduce human interference in network management, the research community has lately begun to investigate AI [43]. The intelligence-defined networks research group was established by the internet engineering task force (IEFT) to explore the application of machine learning in networking. Moreover, an intelligent management framework can supply self-healing, self-optimization, and self-protection abilities for wireless networks by reactively and proactively coping with network problems utilizing AI approaches [42].

## 2.9 Towards sixth-generation (6G) and Motivation

A slew of new apps and use cases inspired by the recent trends in technology are now being developed which strains 5G's capabilities. This has prompted academics to reconsider and develop next-generation mobile communications systems, "6G" [44], [45]. It is predicted that 6G networks will usher in a paradigm shift in mobile networking by achieving extraordinary network capabilities to meet the needs of a data-driven world.

During the past four decades, mobile networks have progressed through five generations. Every ten years, a new development of mobile networks arises, with additional technology and advancements to enable humans to enhance the quality of their life and job experience. Among these generations, 5G mobile networks are still being utilized globally. The eMBB feature in 5G allows for peak traffic speeds of up to 10 Gbps. Moreover, compared to 4G, URLLC reduces

latency to as little as 1 millisecond, while massive machine type communication (mMTC) accommodates over 100 times more devices per unit area. Also, network softwarization is a significant 5G technology that allows networks to be more dynamic, programmable, and abstracted [46].

Edge intelligence (EI), non-orthogonal multiple access (NOMA), communication from sub-6GHz to THz, self sustaining networks (SSN), and large intelligent surfaces (LIS) are just a few of the innovative concepts that have emerged in recent years in the field of communications [47], [48]. Also, predicted to develop as major applications of future communication systems are extended reality (XR), unmanned aerial vehicles (UAV), space, and deep-sea exploration. But nevertheless, the network functionalities promised by 5G do not fulfill the criteria of these applications, which include real-time availability to strong computational resources, ultra-high data rates, incredibly low latency, and exceedingly high availability and reliability [49], [50]. This has prompted researchers to consider 6G mobile communication systems. 6G is planned to make use of new communication technologies that link a wide range of devices, completely support developing apps, and offer users proper access to strong storage and computation capabilities.

# Chapter 3

# 5G Core

## 3.1 Service-Based Architecture

First attempts towards virtualization were started with LTE virtual evolved packet core (vEPC). service-based architecture (SBA) which is used in 5G networks is an evolution of LTE vEPC [51]. Network functions (NFs) are the key components of SBA and the interaction among NFs can be established through two approaches. The first approach, the reference point representation illustrates a point-to-point reference point between the NFs. The reference point diagram is shown in Figure 2.1 as metioned in the architecture section in Chapter 2. In the second method of interaction (SBA) between NFs , the SBI are being deployed between NFs of the control plane as depicted in Figure 3.1. The NFs of 5G control plane is represented in the top part of Figure 3.1. Instead of point-to-point connections, the NFs of the control plane are connected by a network bus. The entity name is contained in the interface name that is prefixed with "N", whereas, the point-to-point interfaces are designated by "N" and a number in the lower section of Figure 3.1 [52]. In SBA a common framework is utilized among various NFs to expose their services to other NFs. Moreover, in SBA, the HTTP-based APIs are utilized instead of conventional telecom signaling methods [53]. In this architecture, network repository function (NRF) acts as a centralized repository for all the NFs. Also, the NRF allows every NF to recognize other NFs services. The network is more flexible under a service-based architecture, and it can

rapidly adapt to unforeseen needs. However, if a new network entity requires to be defined in a point-to-point architecture, multiple new interfaces and protocols should be standardized for connecting to other entities. This frequently results in a complicated network [52].

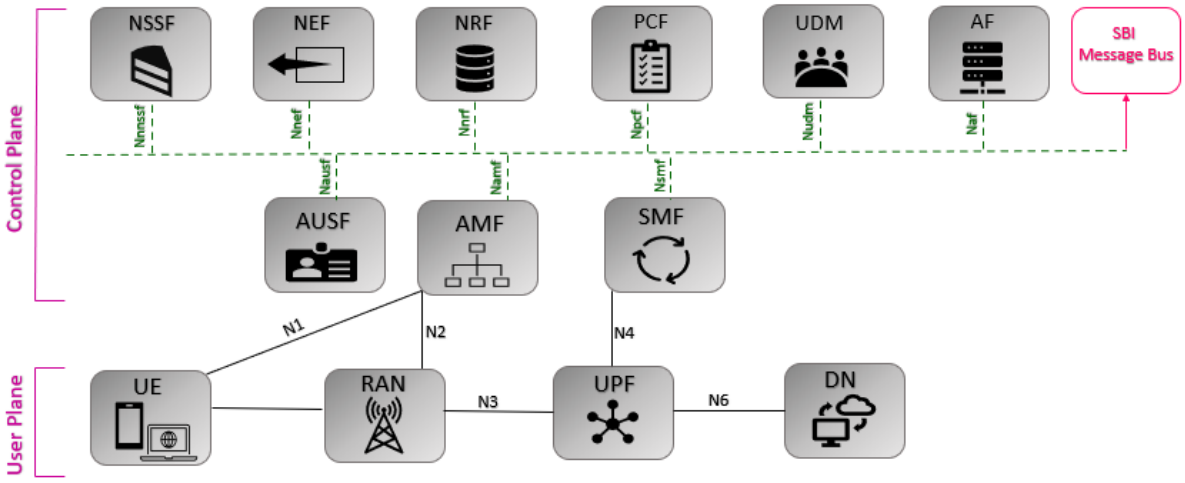


Figure 3.1: Service-Based Architecture

Also, a producer-consumer model is used for interacting among NFs over SBI. Hence, a service provided by an NF (Producer) would be used by another NF (Consumer) who has been authorized to do so. NFs deploy one of the following methods as presented in Figure 3.2 to communicate with each other:

- **Request-response approach**: In this method, a producer NF would provide a service to a consumer NF when it is requested.
- **Subscribe-notify approach**: In the subscribe-notify methodology, a consumer NF subscribes to the services that are provided by the producer NF that would notify the subscriber of the result.

The SBA allows for more decoupled and granular network functions. Also, this architecture allows various services to be upgraded independently with minimum impact on other services. In this architecture, services are on demand basis which enables the automation and reduction of delivery time.

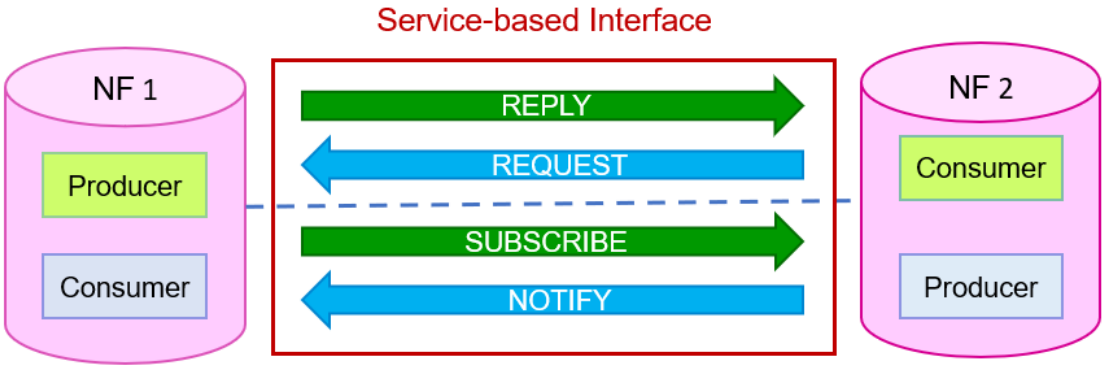


Figure 3.2: NF Service Operations

## 3.2 5G Core Architecture

In the 5G network architecture, network slicing, and switching are leveraged to obtain better efficiency in different scenarios, and each new advancement in this technology acquires higher efficacy, capabilities, and use cases in diverse scenarios [54]. In the development of mobile wireless communications, the 5G core architecture has the goal to have the largest range of services and applications [55]. URLLC, eMBB, and mMTC are the three types of 5G innovative applications. The development objectives for 5G include agility, programmability, dependability, resilience, multi-tenancy compliance, and economical resource usage [56]. To attain the specified objectives, 5G core is based on SBA which is elaborated in previous section and in Figure 3.1. It denotes that the 5G core is made up of separated NFs that are connected and each of which can use the services of the others if that NF is allowed to do so [57]. Since 5G core architecture leverages NFV, SDN, and service-based approaches, its architecture is considered as “cloud-native”. The following are some of the most important 5G core design concepts that should be taken into consideration [58]:

- **control and user plane separation (CUPS)**: CUPS is utilized to provide decoupled technological progress. The control plane is separated from the data flow. However, they are connected together via standardized interfaces which leads to greater flexibility and scalability [58].

- **Decreasing the access network (AN) and core network (CN) dependencies**: Operators are able to design a convergent and multi-access core network using standard AN-CN interfaces that incorporate many 3GPP and non-3GPP access forms [58].
- **Supporting the stateless NFs**: In this concept which is inspired by cloud applications, compute and storage resources are separated. It also makes the usage of network processing pathways much more efficient [58].
- **Exposing the network capabilities**: In 5G, it is crucial to expose the information of the network's capabilities with external and internal applications. This is particularly important when operators attempt to incorporate 5G with vertical industrial operations. Vertical clients, especially those with worldwide operations and multi-operator partnerships, benefit from standardizing the exposure of network capabilities [58].

## 3.3 5G Core Main Network Functions

Accessibility and communication between services are enabled via network functions. Every NF is able to behave as both a producer and a consumer of service when a service is required. Some NFs are associated with the control plane, whereas others are related to the user plane. As the network must always perform some basic functions including communication with the UE, storing its subscription, permitting the access to external services and networks, controlling access and mobility, and providing and maintaining security, some of NFs possess a high degree of similarity to the CN of the earlier generations. However, several NFs that have not been introduced previously are added to the 5G core architecture for enabling new network concepts such as service-based interactions and slicing. NFs are software-based which allows them to be adjusted according to necessity, for assisting the enablement of various data services and requirements. The 5G core is made up of the following significant NFs that are elaborated below [59]:

- **AMF**: AMF has a similar role to the mobility management entity (MME) in LTE. The AMF also plays a key role in authenticating the subscriber within the network, and pro-

vides the device with a temporary identity which can be used whenever it signals the network. The temporary ID is used in paging as well. The AMF can also [59] terminate non-radio signalling (NAS), protect NAS ciphering and integrity, monitor registrations, connections, access authentication, and authorization.

- **SMF**: The next element in 5G core is the SMF. Traditionally in LTE, the MME would perform mobility management of AMF and session management of SMF. In 5G, these functionalities have been split between AMF and SMF. The SMF is directly involved in protocol data unit (PDU) session establishment and modification. It will be routinely used with the policy control function to determine whether or not a particular user data session is allowed to proceed.

- **UPF**: In this proposed 5G core architecture, UPF is the only NF related to the user plane. The UPF is basically an integration of SGW and PGW, which are the data plane sections in the 4G LTE. The UPF is responsible for conducting the routing and forwarding user data packets between the gNB and the external WAN. It also has a connection to the SMF. UPF is an anchor point for the next generation radio access network (NG-RAN) mobility. Moreover, UPF ensures that the right data is sent down the correct QoS flow and implements an appropriate policy.

- **PCF**: Next in line is PCF, which is used to implement the policy control function and enforce the subscriber policies. The implementation of policy control is on a dynamic basis and these dynamic policy decisions are based on conditions that might be active in the network.

- **AUSF**: As its name represents, AUSF is an authentication server that decides on the authentication of UE.

- **UDM**: The UDM can generate authentication and key agreement (AKA) credentials, access authorization, handle user identification, and monitor subscriptions.

- **AF**: The AF allows applications to affect traffic routing, access network exposure function

(NEF), and communicate with a policy framework.

- **NEF**: NEF facilitates the exposure of capabilities and events, as well as the secure transmission of data from an external application to the 3GPP network and the translation of external and internal data.

- **NSSF**: NSSF redirects traffic to a network slice. In addition, the NSSF executes the selection of the network slice instance for serving the UE and determines the NSSAI and the AMF to be employed to serve the UE.

- **NRF**: Network Functions can recognize each other thanks to the NRF's service registration and discovery feature. It also keeps track of the NF profile and available NF instances.

## 3.4 Network Slices

Network slicing, an important method utilized by the 5G network, is a research priority in both academia and industry. The next generation mobile network (NGMN) Alliance describes network slicing in the scope of 5G as a technique that allows the incorporation of physical and logical network and cloud resources into the accessible, software-based, multi-tenant, programmable, network environment [6]. It entails connecting multiple self-contained logical networks to a shared physical infrastructure platform, resulting in the creation of a flexible ecosystem of stakeholders that fosters technological and business development.
A network slice is a separate network that contains its own virtualized resources, traffic flow, and topology. The 5G core network consists of four slice types including default slice, eMBB, URLLC, and mMTC which are explained below (Figure 3.3):

- **Default slice**: To facilitate UE registration to the 5G core network for all service slices, the default slice consists of widely used network functions such as AMF, NSSF, PCF, UDM, and AUSF. The default slice authenticates that the UE is permitted for network access when it initially connects to the 5G core. Moreover, the default slice maps service types that the UE brings into matching slice IDs that the UE will utilize, and allocates

a temporary ID and IP addresses to the UE. Policies such as load balance and regional distribution can be used to install the default slice. Each default slice is given its own slice ID in this circumstance [60].

- **eMBB**: eMBB is capable of maintaining stable connections with extremely high peak data rate and is commonly utilised for entertainment purposes, such as event streaming. The applications of eMBB use a wide range of bandwidth. Also, eMBB generates the most traffic of the mobile network.

- **URLLC**: The URLLC is utilized for mission-critical networks. URLLC allows for low-latency, high-reliability transmissions of small payloads from several terminals that are activated in response to external events like alarms [27].

- **mMTC**: mMTC supports slices that contains a high number of IoT devices. Data volumes are often low in this type, therefore traffic related to mMTC does not require a huge bandwidth. Offering high coverage of transfer rate with low expenses is one of the unique features of the mMTC slice.

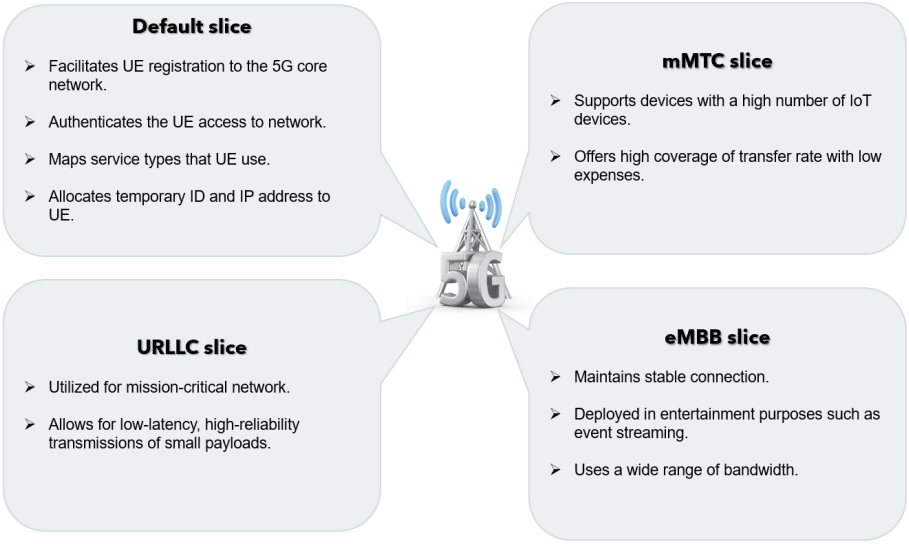


Figure 3.3: Slice Types in 5G

The slices can be fine-granular in the user and service layer or they can be coarse-granular in industry or corporate level such as the connected car slice. The slice should preferably reach

all the way to the UE and run end-to-end all over the RAN, transport, and core domains.

## 3.5 5G Core Deployment Review

For 5G networks, there are two major deployment methods: non-standalone (NSA) or as a standalone (SA).

### 3.5.1 5G NSA

By integrating a 5G RAN with the LTE evolved packet core (EPC), clients are provided with faster data transmission speeds in the 5G network. This is known as a Non-Standalone Architecture because the 5G RAN is still dependent on the 4G core network to handle control and signal information, and the 4G RAN is still viable. Carriers can deliver faster and more reliable eMBB without fully rebuilding their core network technology by employing the current infrastructure of a 4G network [61].

### 3.5.2 5G SA

The operation of standalone 5G is not reliant on an LTE EPC. Instead, 5G radios are incorporated with a 5G core network that is cloud-native. In 5G SA deployment, the 5G core is based on SBA for providing the entire range of 5G functionalities that are required by enterprises for deploying automation in industry, automobiles, and so on [61]. The end-to-end support for 5G services and faster speed are the clear benefits of the 5G SA deployment.

### 3.5.3 Comparison between 5G SA and 5G NSA

There are some advantages of 5G NSA versus 5G SA. For operators in the early stages of 5G rollout, 5G NSA is preferable since by deploying the current 4G infrastructure, the time that takes to launch the network is decreased. This enables telecom operators to deploy their 5G network and deliver 5G services considerably faster than they could if they employed the SA strategy. The cost of network implementation is also lessened by utilizing existing infrastruc-

ture. It alleviates the strain on operators, allowing them to focus on providing high-quality service to clients. However, the advantages of faster deployment and lower costs come at the expense of network performance. Even though 5G NSA is a substantial upgrade over 4G LTE, it is not capable of delivering the full benefits of 5G [62].

The possibilities for 5G SA, which are based on a new 5G network core, are enormous. For instance, ultra-reliable low-latency functionality for fields, such as autonomous devices and next-generation IoT, and massive machine-to-machine communication solutions, will be available with this version of 5G. Also, 5G SA will provide extra bandwidth and lower latency. This type of deployment is required for numerous applications and use cases, including virtual reality and industrial IoT. Figure 3.4 illustrates the NSA and SA deployments of 5G network.

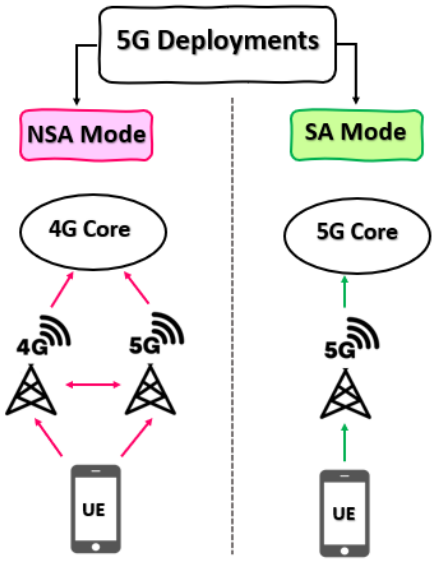


Figure 3.4: NSA and SA Deployment Methods in 5G

## 3.6 4G Core to 5G Core Migration

5G can be implemented in a standalone mode using the 5GC or in a non-standalone mode employing the EPC, as explained in the preceding section. For most operators, properly managing the migration from the 4G core to the 5G core as well as optimizing expenditure are vital parts of a 5GC strategy. A migration [63] from evolved packet system (EPS) to standalone

5G is performed by utilizing inter-RAT mobility methods to shift devices from 4G LTE with EPC coverage to 5G new radio (NR) with 5GC coverage, in this proposed structure. One of the main advantages of this approach is that SA architecture can fully leverage 5G end-to-end capabilities provided by NR and 5GC, allowing for the efficient supply of customised services, particularly to vertical industries. This migration results in a more enhanced and customized user experience by enabling new capabilities such as network slicing, SBA, and MEC based on the certain requirements of each service [63].

Interworking between the 5GC and the current EPC is the ideal way to enable services demanding wide-area coverage during the migration phase when NR coverage is being established. As user data and policies must cover two networks (the EPC and 5GC), interworking with the EPC puts a strain on the backend business support system's (BSS) integration. New devices must have 5G capabilities, whereas devices that merely support the EPS, such as incoming roaming devices, will remain for a long time and will need network maintenance. This long-time requirement is a compelling justification for a dual-mode core network solution that incorporates both EPC and 5GC. The defined operating paradigm for the EPC and 5GC, which simplifies overall system supervision, is a fundamental advantage of a dual-mode core network solution [64].

### 3.6.1 The architecture of interworking between 5G and EPS

Presenting the 5G to a network demands a complete plan that takes into account all network domains, spectrum assets, coverage techniques, and gadgets and also determines where each service should be provided. New NFs and interfaces are presented via 5G internally and through operation support systems and BSS such as charging systems. As for the NG-RAN, 5G also contains innovative protocols and interfaces, which means there must be coordination between the RAN migration and the 5GC rollout. Furthermore, SBA containing a network repository function for registering and discovering a service along with new features such as network exposure and network slicing should be addressed while introducing the 5GC [64].

Operators possessing both NR and LTE access can leverage 5GC capabilities for tight inter-

working to the EPS, which was formerly referred to as EPC-5GC tight interworking in the 3GPP's initial release of 5G standards [65]. The 5G architecture for EPC-5GC tight interworking is illustrated in Figure 3.5. In this figure, purple lines indicate signaling while the blue lines represent the user plane interactions. The 5GC architecture comprises a common user plane (UP) anchor point recognized by the session management function with the packet DN gateway control plane function (SMF+PGW-C) and the user plane function along with the PGW user plane function (UPF+PGW-U) to enable IP address preservation while connecting over and shifting between 4G and 5G access. Moreover, the MME and the AMF communicate directly through the N26 reference point, which supports devices in single-registration mode, to provide smooth service continuity and network-controlled handover. In fact, the device is either registered in the MME or the AMF, but not both at the same time. In addition, mapping the protocol data unit (PDU) sessions in the 5G to packet data network (PDN) connections in the EPS and vice versa is part of tight interworking.

The interworking architecture guarantees that new 5GC-capable devices are constantly connected to the UPF in the 5GC, regardless of whether they are connected via 4G or 5G access, preserving IP addresses when devices switch accesses. For a device with NR or LTE access, policy and subscription management must be offered in a stable manner. Furthermore, through EPC, the interworking architecture supports numerous 5GC features, including network slicing capability. By transferring the packet gateway, subscription, and data management, providing additional features, and policy control, operators can enable the migration from a single-mode EPC to a dual-mode EPC and 5GC network solution.

### 3.6.2 The scope of packet core

The packet core scope contains capability for MME, AMF, the serving-gateway-control plane function SGW-C, PGW-C, SMF, and UP functionality (SGW-U/PGW-U, UPF), and session management. The 5GC was first introduced for wide-area services, allowing RAN and core migration to occur independently of one another. Moreover, a split of the gateway functions in the EPC into control plane (CP) and UP, defined as CUPS, was standardized by 3GPP previous

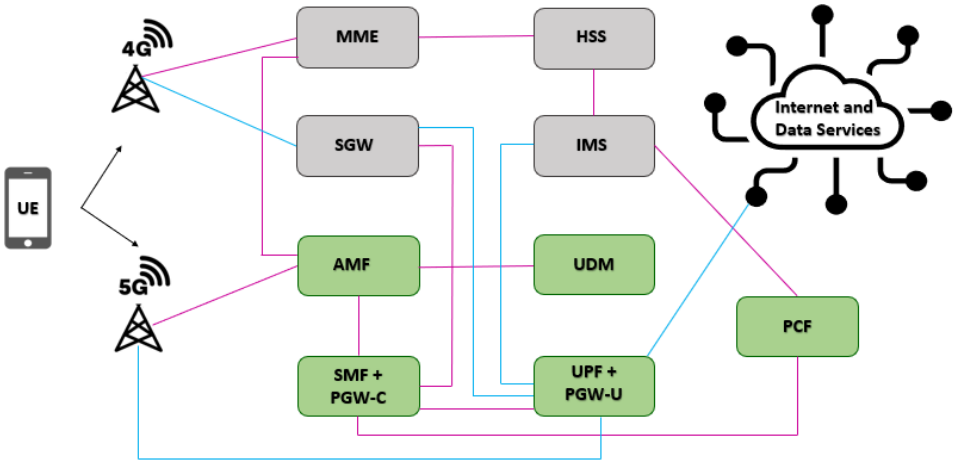


Figure 3.5: The Architecture for EPC-5GC Tight Interworking

to the standardization release of 5G. CUPS opens up new possibilities for UP distribution and edge breakout of existing EPC traffic. In central and local deployments, for instance, separate PDN connections can deploy various SGW-U/PGW-Us. Furthermore, separating the CP and UP functions in the 5GC architecture is performed over the SMF and UPF respectively.

Several migration pathways from the EPC to the 5GC are feasible due to the existence of both CUPS and EPC-5GC tight interworking. Before migrating to the 5G, one possibility is to first incorporate CUPS into the EPS, allowing the operator to leverage the CP and UP segregation. This approach would be effective in dealing with growing traffic demand. In terms of NR NSA deployment and plan for a seamless migration to the 5GC based on a UP implementation that supports both the EPC and the 5GC, CUPS would be advantageous.

Another approach involves integrating SMF+PGW-C with SGW-C functionality and UPF+PGW-U with SGW-U functionality for implementing CUPS simultaneously as the 5GC. While connecting through either 4G or 5G access, the new high-capacity 5G devices can be handled by a CP and UP split-gateway design, as indicated in the middle section of Figure 3.6. The ability to dispense the UP more flexibly in multiple places is a further perk of this method. For instance, if we put the UP adjacent to the RAN, low-latency services can be assessed. The center section of Figure 3.6 further demonstrates that when deploying the 5GC, older devices with 4G-only subscriptions may continue to employ the current SGW and PGW functions in the EPC, minimizing the effect on existing customers and services. Also, for gadgets with 5GC subscriptions

and 5GC-NAS (non-access stratum), the gateway selection (SGW-C & SMF+PGW-C) capabilities must be handled by the MME. Also, more approaches including dedicated access point names or domain name system lookup enhancements, can be deployed by the MME to facilitate the gateway selection.

Merging the whole EPC and 5GC capability into a dual-mode packet core, which will serve 5G-enabled subscribers, is the next migration phase as indicated in the bottom side of Figure 3.6. In this scenario, the SMF+PGW-C is used to service the new 5GC devices, whereas a separate PGW instance is utilized to serve the traditional 4G-only devices. Other deployment models are possible as well. The migration's objective is to provide a solution that adheres to the concepts of a common operational model in accordance with cloud-native deployment and helps both new 5G-enabled customers, irrespective of access technology, and legacy 4G customers, who are only connected via the EPC.

## 3.7 Role of AI in 5G

AI is the application of scientific knowledge which makes machines as clever as humans, and it has long been utilized to improve communications systems in a variety of configurations [66]. Significant advancements in AI and computation have motivated communication researchers to deploy AI in the 5G network. For instance, the establishment of an intelligent and complete data repository can be performed by an AI-defined 5G network via separating, analyzing, and interpreting operational data.

Furthermore, AI algorithms are also utilized to tackle security concerns. In fact, the information security sector is creating an increasing amount of data, which exposes them to advanced attacks, and AI is currently used as an effective countermeasure. Indeed, threat detection, data analysis, and human assistance are performed by the first generation of AI, whereas less human-dependent systems which operate autonomously are handled by the second generation of AI [67].

The 5G network proposes a wide range of technical solutions, including increased band-

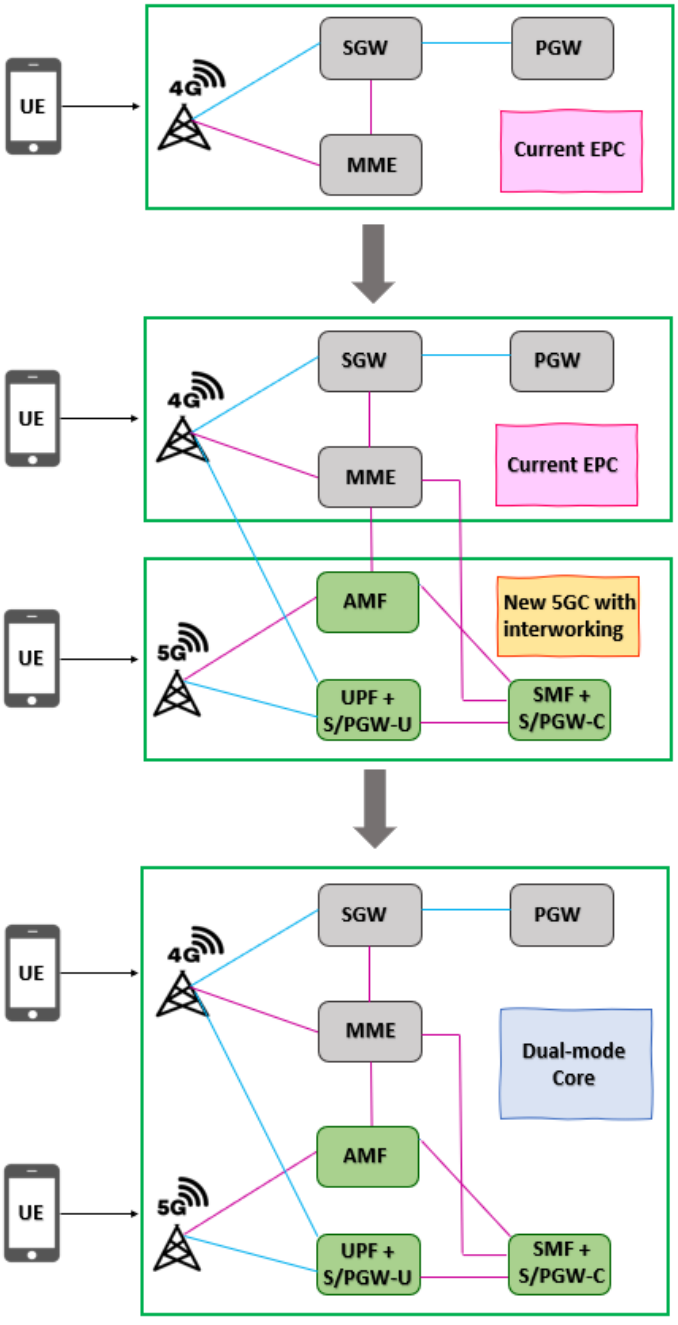


Figure 3.6: Migration from 4G to 5G Utilizing EPC-5GC Tight Interworking

width, continuous connectivity, reduced latency, large data capacity, and enhanced quality of service (QoS). AI is one of the significant technologies that is currently used in 5G. It supports a wide range of aspects related to the technological requirements [68]. Also, machine learning (ML) is a subset of statistics that entails generating techniques which allow computers to learn from statistical patterns in data and create an analytical model without having to be manually programmed [69]. Therefore, devices which utilize machine learning capabilities can learn from data, recognize data patterns and make decisions with little human intervention.

### 3.7.1 The application of ML in 5G

The use of ML in the field of 5G has caught the interest all around the world. ML has the potential to play a critical role in network automation, lower operational costs, and enhance customer experience. 5G network modeling, optimization of a non-linear network objective function [70], parameter evaluation based on prior experience, anomaly detection, and fraud prediction, all utilize ML.

### 3.7.2 Learning approaches

The discriminative characteristics of a system that cannot be elaborated by mathematical principles are learned via ML models. A model which has been trained on the supplied data, can make decisions on unseen data based on the learned patterns and execute tasks employing arithmetic calculations. This would enable ML modeling to be accessible and portable based on data. The automation in network management can be facilitated by utilizing ML approaches. For instance, the key performance indicators (KPIs) within predefined thresholds help to maintain the performance of network management. Also, ML models can improve troubleshooting, energy usage, and QoS. ML comprises three main subcategories, which are described below:

- **Supervised learning:** Supervised learning (SL) employs labeled datasets to conduct a mapping from the input data to the known output targets. SL deals with two main tasks of regression and classification. Decision trees (DTs), k-nearest neighbours (KNN), and Gaussian Process Regression (GPR) algorithms are some examples of SL approaches.

- **Unsupervised learning**: Due to the fact that data labeling is not always feasible, unsupervised learning (uSL) has been introduced to discover underlying data patterns and associations between unlabeled input data values and cluster them. K-means clustering and principal component analysis (PCA) are two examples of unsupervised learning approaches.

- **Reinforcement learning:** Reinforcement learning is a technique in which the data is not predefined and a learning agent interacts with its surroundings and learns to map each input to a certain action and determining the output by trial and error. Q-Learning, policy learning, and the Markov decision process (MDP) are examples of Reinforcement learning approaches [71].

## 3.8 5G Core, Edge Computing and Next Generation RAN

Latency constraints between the UE and the computing/storage platform are becoming increasingly rigorous in the 5G era. Traditional cloud computing may not be able to meet these new latency requirements, as ultra-low latency has become one of the major aspects of 5G technology. To meet these needs, a new paradigm is needed. The European Telecommunications Standards Institute (ETSI) proposed the mobile edge computing paradigm, which was renamed as multi-access edge computing (MEC) in 2017 [72]. In fact, MEC's main idea is to move cloud computing storage and processing functions to the mobile network's edge.

Operators have an impact on the sites of base station controller and transport aggregation. There is a possibility to turn the centralized data centers into distributed ones. These data centres may be utilised to terminate access connections from the 5G RAN and become the apparent area to implement 5G core functions, particularly user plane functions. Furthermore, they can host latency-sensitive applications, based on service needs. In fact, a distributed computing system in which data processing takes place near the edge where data is generated is called Edge computing [73].

By leveraging edge computing, the storage and processing of data and computing capabilities

are located nearer to the gadgets and devices which generate the data. Some objectives of leveraging edge computing include decreasing the latency between serving applications and UE and offering the bandwidth that is needed for the core network. Along with delay enhancement, there are other benefits in utilizing MEC in 5G. Some of these advantages can be included as expense savings and optimzed network usage through employing shorter data traffic paths. Also, computation and network resources would be managed in NFV scope. Most operators are undergoing a lengthy network transition in preparation for 5G which involves a transition to full virtualization and MEC is an important step in this process.

The content delivery networks (CDNs), which provide video material to consumers from edge servers located near them, rolled out the edge computing. These edge servers subsequently evolved and began to host applications [74] leading to an innovative distributed framework currently known as edge computing. Centralized data centers cannot assure the necessary latencies and transfer data rates due to the considerable growth of devices in the network edge. High service quality criteria such as ultra-low service latency that cannot be met by cloud computing can be achieved by utilizing the edge computing. Since the 3GPP architecture integrates fixed access networks (FAN) and RAN, edge computing is not just associated with cellular networks. MEC, like edge computing, minimises service latency and bandwidth consumption by locating MEC hosts which process user data near to the end-users. While MEC applications are correctly deployed in these MEC hosts, hosts can be located with base stations and other cellular nodes that are more adjacent to the network edge, such as radio network controllers (RNC) and base stations. MEC is significant in 5G because it can assist in creating a service environment with ultra-low latency and real-time access to radio network information, both of which are key properties of the technology.

Finally, a high-level architecture of MEC in 5G is illustrated in Figure 3.7. With the support of the MEC system, an edge application is located near the RAN and user plane in this architecture. While the user accesses the application on the user equipment, depending on the user location, the MEC system leads the 5G core network to choose a user plane function for generating a traffic session closer to the user location. UE gets access to the application from

the edge based on the edge application and user plane selection which decreases the latency.

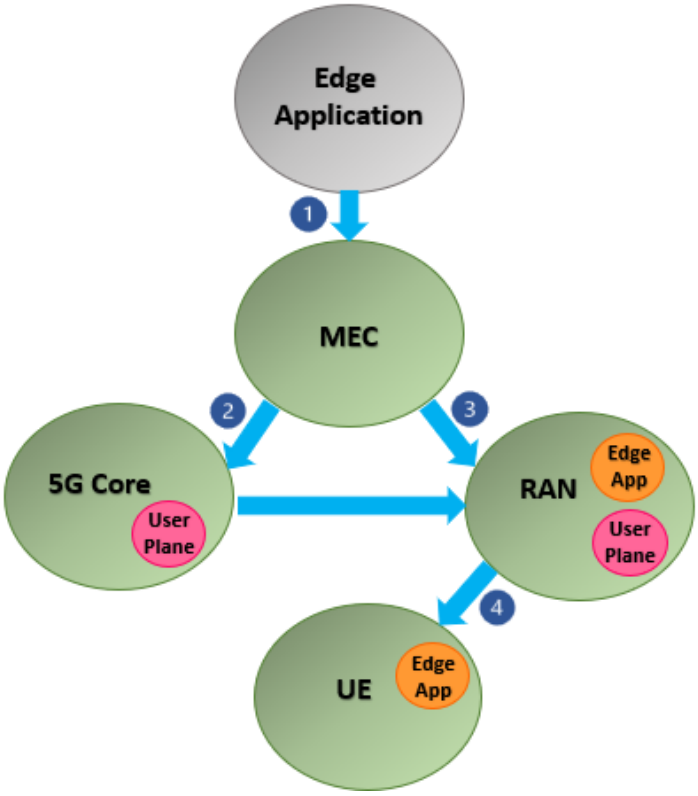


Figure 3.7: MEC Architecture in 5G

# Chapter 4

# Application of Machine learning to Traffic Forecasting in 5G Networks

The notion of a network slice, which is described as a virtualized subset of the physical resources of the 5G infrastructure, is used in the standard for 5G communication. It is vital to assess the resource requirement of each network slice and how it evolves over time since a large number of network slices are deployed over a 5G network. This enhances the resource efficiency without compromising network slice performance as a typical method of determining resource demand in traffic forecast. The efficiency of ML predictors for traffic prediction in 5G networks has been established in cutting-edge research. A deep-learning based analysis of a traffic dataset is conducted in this chapter. In fact, temporal and spatiotemporal models were experimented. In the temporal analysis, the study explores the forecasting performance of the FCSN, 1D-CNN, SS-LSTM, and AR-LSTM models. Moreover, to predict the next 24 hours of cellular traffic and incorporating the spatial and temporal dependencies, a 2-dimensional Convolutional LSTM model (2D-ConvLSTM) is developed in a multi-channel framework. Furthermore, baseline models for both frameworks are presented to assess the performance of the aforementioned models.The results reveal that FCSN and 1D-CNN have comparable performance. However 1D-CNN is a smaller network with less number of parameters. One of the other benefits of the proposed 1D-CNN is having less complexity and faster execution time for

predicting the next 24-hour traffic.

Also, the proposed 2D-ConvLSTM model outperformed the spatiotemporal baseline model in terms of MAE and RMSE metrics for forecasting all three types of cellular traffic including internet, sms, and call. The high performance of the 2D-ConvLSTM model is due to its strong ability in capturing spatiotemporal dependencies simultaneously. Also, we shrank the network via progressive channel sorting for optimization to reduce the number of parameters.

## 4.1 Introduction

Mobile devices with Internet access are infiltrating every part of people's lives, including their job and leisure. The proliferation of smartphones, as well as the advent of increasingly diversified applications, have increased cellular data traffic. Due to the shift in customer preference for wireless access, present mobile infrastructure is experiencing significant capacity constraints. To meet this rising need, pioneering approaches propose deploying resources more quickly and tackling mobility management in a distributed manner [75], [76] [77]. Nonetheless, in the long term, intelligent heterogeneous architectures and technologies capable of spawning the 5G, should be built to address more demanding end-user application requirements [78], [79] [80].

5G employs a variety of technologies to satisfy these requirements, with network slicing being one of the most important. In fact, the splitting of a physical network into numerous virtual networks, each of which may be modified and optimized for a certain kind of application, is referred to as network slicing. The shared physical network resources may be dynamically and effectively allocated to logical network slices depending on changing user needs by leveraging cloud computing and virtualization technologies. It is essential to identify the resource requirements for each slice and how these requirements change over time. If a network slice requires more resources than those initially allocated, it is considered under-provisioned. This results in poor network slice performance and QoS for users [81]. In contrast, if the network slice utilizes fewer resources, it is over-provisioned. In this case, resources are not required but active. this

incurs expenditures on infrastructure providers. Dynamically adjusting the resources allocated to network slices is crucial as both of these scenarios (over- and under- provisioning) cause expenses on infrastructure providers and leads to quality of service deterioration [82]. Therefore, recognizing the traffic profiles of each slice is critical for allocating resources dynamically.

Also, communication networks are becoming more intelligent and self-organized as the 5G technology evolves [83]. The SON must adapt to changing usage patterns and take proactive measures. As a result, using big data to anticipate and analyze mobile traffic, provides the cornerstone for smart management features and is extremely important in the industry [84].

ML and time series analysis, which have been used in a variety of applications, are considered as powerful tools for modelling and forecasting network traffic. Incorporating adaptable machine intelligence into the future mobile networks is attracting a lot of attention in the scientific community [85] [86][32]. This trend is represented in the development of networked systems that leverage ML techniques to tackle challenges ranging from radio access technology (RAT) selection to malware diagnosis [87],[88]. ML allows for the systematic extraction of useful information from traffic data and the automated discovery of correlations that otherwise is too complicated for human specialists to obtain.

In this study, the prediction performance of several neural network models are evaluated in forecasting the mobile traffic. Specifically, the temporal and spatiotemporal frameworks are developed for a univariate and multivariate analysis to predict the next 24 hours of cellular traffic. This chapter is structured as follows. In section 4.2, related works are presented. Section 4.3 describes the details of the analyzed dataset. In Section 4.4, the applied neural network models are introduced. The experimental results are provided in Section 4.5. Section 4.6 depicts discussion and comparison.

## 4.2 Related works

Data created by mobile devices is extremely varied since it is often collected from multiple sources and has various formats [89]. Traditional machine learning methods become infeasible

to solve a variety of challenges within this scope such as data abstraction and extraction of a meaningful embedding. Also, when the data scale increases, the performance does not enhance [90], and in control problems, they cannot manage highly dimensional spaces [91]. Therefore, the incorporation of deep learning in 5G mobile and wireless networks is highly justified with an automated and intelligent data representation and feature selection. This implies that data can be effectively distilled and higher abstractions can be recognized while decreasing the requirements for the pre-processing. Recently, deep learning-based algorithms have been explored to find potential representations of Internet traffic flows. Oliveira *et al.* [92] investigated internet traffic prediction techniques based on the recurrent neural network (RNN). The proposed RNN was found to outperform the stacked auto-encoders in network traffic prediction. Wang *et al.* [93] proposed an approach that combined an auto-encoder with a long short term memory (LSTM) network to take advantage of distinct cells' spatial dependency. However, the representation learnt by auto-encoder was a lossy depiction of the original data [94], and they may not adequately capture the spatial dependency of nearby cells. In addition, Zhang *et al.* [95] presented a new technique for citywide traffic forecast that takes advantage of the tremendous capabilities of a deep convolutional neural network (CNN). More precisely, the densely connected CNN [12], was used to represent the spatial and temporal dependency of traffic in distinct cells collectively. The convolution operation naturally captures the spatial dependency along with two CNNs which are used to model two temporal dependencies named proximity and period. However, one of the challenges of the proposed approach is the requirement for a large training dataset.

Recently, Lin *et al.* [96] suggested an intelligent data-driven BS sleeping mechanism that analyzes the capacity of BSs in various locations. A prediction model for spatio-temporal cellular traffic was proposed with a multi-graph convolutional network (MGCN) for capturing the spatial information. Also, the temporal features are extracted using a multi-channel LSTM system that includes hourly, daily, and weekly periodic data. The proposed MGCN-LSTM model performed better compared to other models in terms of cellular traffic forecast accuracy. Furthermore, Chien *et al.* [97] explored spatio-temporal dependencies among base stations and

suggested a data preprocessing method based on CNN architecture to extract spatio-temporal features. In comparison to previous deep learning methods, their proposed strategy can reduce neural network parameters and requires just a minimal amount of processing cost to estimate cellular traffic. To investigate spatial and temporal sequence information at the same time, Liu *et al.* [98] devised a Spatial-Temporal Transformer (ST-Tran). The ST-Tran can simulate real-time spatial correlations between grids in a global area by analysing cellular traffic data of all grids during one specified time interval as a spatial sequence. Also, regarding the energy-saving approaches, Gao *et al.* [99] presented two load prediction models to anticipate traffic load in cells. A linear ensemble model made up of three sub-models, was one of the proposed methods for predicting traffic load. Various approaches such as linear regression, and regression tree were applied to the sub-models. A residual CNN (ResNet) was utilized to train the collected data. The ensemble model outperformed other baseline models in terms of prediction accuracy, while ResNet enhanced calculation efficiency [99]. To deal with the task of large-scale traffic prediction, Zhou *et al.* [100] suggested an attention mechanism to create a new spatial-temporal graph convolutional network (STA-GCN). This research presented a regional transfer learning technique based on STA-GCN in order to accomplish large-scale traffic prediction. The experiments indicate that a transfer learning technique may successfully increase knowledge reusability and speed up model fitting without sacrificing prediction accuracy. However, to capture spatial dependencies effectively, dynamic graphs are required. Moreover, for anticipating mobile cellular traffic, Zhao *et al.* [101] suggested a spatial-temporal aggregation graph convolution network (STAGCN) to extract the complicated spatial-temporal characteristics at various timestamps. The external elements were then merged with the outputs of the aggregating GCN modules to acquire the final anticipated traffic. Also, a dual-channel-based graph convolutional network (DC-STGCN) model was introduced by Pan *et al.* [102] where two temporal components represent the network traffic correlation on a daily and weekly basis. A gated recurrent unit (GRU) in each of the two components extracted the spatial-temporal characteristics. The correlation and adjacency feature extraction modules were also included in the model to capture node connectivity and proximity correlation, respectively. The GRU also captures

the traffic's temporal properties. The suggested model outperformed current models and could make long-term forecasts, according to their results. Furthermore, to successfully capture the intricate patterns concealed in cellular data, Zhang *et al.* [103] presented a new deep learning architecture called STCNet. This research collects three types of cross-domain datasets, namely, BS information, point of interest (POI) distribution, and social activity level, in order to produce a complete characterization of external variables that impact cellular traffic volume. To enhance cellular traffic prediction, the correlations among these datasets and internet traffic prediction were analyzed. Also, to anticipate cellular traffic, Zhao *et al.* [104] suggested a spatial-temporal attention-convolution neural network to capture the daily and weekly temporal dependencies of traffic data and external factors. Finally, the Spatio-temporal cross-domain neural network was presented by Zeng *et al.* [105] (STC-N) and the influence of varying amounts of cross-domain big data on traffic forecast accuracy was explored. Furthermore, the authors fused transfer learning across services and between classes to create a new technique called Fusion-transfer. According to their experimental results, the model outperformed the No-transfer and Part-transfer models.

## 4.3 Dataset

A multi-source dataset generated by Telecom Italia in 2015 [3] is used in this study. This open-source dataset is among the most complete collections from an operator. The collection was initially developed for addressing a big data challenge with concepts spanning from mobile networking to social applications. This dataset is made up of data records regarding telecommunications, weather, news, social networks, and electricity for the city of Milan and the Province of Trento from November 1, 2013 until January 1, 2014. We used the telecommunication records from Milan to predict the traffic in this study. Data recording begins with the creation of geographical grids which divides the city into 100×100 sections using aggregated call detail record (CDR) data. Each grid has its own square ID and covers an area of 235 x 235 meters. Table 4.1 shows the information which was utilized in this study from the

telecommunications records.

Table 4.1: Features of telecommunications records.

| Features | Description |
|---|---|
| **Square ID** | The identification of the Milan Grid square. The square ID and the grid ID are identical concepts that are used in this study. |
| **Time Interval** | 10-minute time interval from the beginning of the record's time. |
| **SMS-in Activity** | Number of SMSs received inside a certain square id and over a specified time interval. |
| **SMS-out Activity** | Number of SMSs sent inside a certain square id and over a specified time interval. |
| **Call-in Activity** | Number of calls received inside a certain square id and over a specified time interval. |
| **Call-out Activity** | Number of issued calls inside a certain square id and over a specified time interval. |
| **Internet Traffic Activity (MB)** | Number of created CDRs in this square id throughout the time span. |

## 4.4 Methodology

### 4.4.1 Problem definition

In this study, our purpose is to predict telecommunication traffic by using historical data. Specifically, based on the previous 24-hour data, we aim to predict the next 24-hour telecommunication traffic including the traffic of sms, calls, internet, count, and frequency features. In this respect, the temporal and spatiotemporal frameworks are developed to predict traffic from the univariate and multivariate perspectives. In the temporal framework, the FCSN, 1D-CNN, SS-LSTM, and AR-LSTM are utilized to predict next 24 hours cellular traffic in the sms, call and internet time series individually. In order to incorporate the spatial and temporal data dependencies a multivariate analysis is conducted as well. For the spatiotemporal framework, a 2-dimensional Convolutional LSTM model is proposed to predict the next 24 hours of cellular traffic using the multi-channel data including sms, call, internet, and count. Furthermore, the baseline models are presented to evaluate the performance of the described models for both frameworks.

### 4.4.2 Preprocessing

The required data for analysis was obtained by grouping the original dataset corresponding to "datetime" and "squareid". For each Grid ID, one subset was created (totally, 10000 subsets). Also, a major amount of cell traffic is zero within the 10-minute time frame of the dataset, which makes the data highly sparse. Furthermore, resource planning at the 10-minute level is a difficult task that might lead to an unstable network or extreme overhead. Therefore, data was re-sampled hourly by summing the traffic. We also added three new measures,"count", "sms" and "calls" to the dataset, where "count" illustrates the number of records in the specific time for a particular grid id, "sms" shows sum of "smsin" and "smsout", and "calls" represent sum of "callsin" and "callsout" because the total traffic of "sms" and "calls" are important.

#### 4.4.2.1 Feature extraction

We extracted days of the week from the "datetime" column. Weekends were separated from the weekdays which means Saturdays and Sundays are indicated by True while the weekdays by False in the "weekend" column. Next, using the holidays in November and December, and the first day of January and generated a "holiday" feature. Also, a new feature named "part_of_day" in which the hours between 6 am and 6 pm are considered as day-time indicated by 1 and night-time by 0.

#### 4.4.2.2 Correlation analysis

Subsequently, the correlations among different variables were calculated, as shown in Figure 4.1, with a correlation coefficient in the range of -1 to 1. A coefficient close to 1 indicates a significant and positive relationship between the two variables, implying that as one grows, the other will be increased, and if one decreases, the other will be reduced as well. A coefficient around -1 shows a significant negative relationship between the two variables, indicating that observations with a large value in one variable are likely to have a lower value in the other, or vice versa. Moreover, there is no linear relationship between the two variables if the coefficient is close to zero. From Figure 4.1, we can see that there is a high positive correlation between

"sms", "internet", and "calls". Also, there is a slight correlation between the "part_of_day", "count","internet", "sms", and "calls".

| | squareid | internet | count | sms | calls | weekend | holiday | part_of_day |
|---|---|---|---|---|---|---|---|---|
| **squareid** | 1.000000 | 0.134045 | 0.111384 | 0.114984 | 0.110346 | -0.000003 | 0.000103 | -0.000087 |
| **internet** | 0.134045 | 1.000000 | 0.442599 | 0.872689 | 0.835118 | -0.034976 | -0.043935 | 0.100908 |
| **count** | 0.111384 | 0.442599 | 1.000000 | 0.490132 | 0.512058 | -0.114089 | -0.039152 | 0.253803 |
| **sms** | 0.114984 | 0.872689 | 0.490132 | 1.000000 | 0.917543 | -0.070013 | -0.020092 | 0.134213 |
| **calls** | 0.110346 | 0.835118 | 0.512058 | 0.917543 | 1.000000 | -0.088326 | -0.052294 | 0.136981 |
| **weekend** | -0.000003 | -0.034976 | -0.114089 | -0.070013 | -0.088326 | 1.000000 | -0.082098 | 0.000410 |
| **holiday** | 0.000103 | -0.043935 | -0.039152 | -0.020092 | -0.052294 | -0.082098 | 1.000000 | 0.010747 |
| **part_of_day** | -0.000087 | 0.100908 | 0.253803 | 0.134213 | 0.136981 | 0.000410 | 0.010747 | 1.000000 |

Figure 4.1: Correlation among variables

#### 4.4.2.3 Statistical analysis

Also, the skewness and kurtosis of all features are calculated as shown in Table 4.2. Since the skewness of all features is positive, features have long right tails. Also, as the kurtosis of internet, count, call, and sms are greater than 3, this indicates that the dataset has a heavier tail than the normal distribution which is illustrated in Figure 4.2.

Table 4.2: Skewness and kurtosis of features

| Feature | Skewness | Kurtosis |
|---|---|---|
| **Internet** | 7.37 | 102.40 |
| **Count** | 1.78 | 6.06 |
| **Call** | 7.96 | 111.51 |
| **SMS** | 8.97 | 117.78 |

#### 4.4.2.4 Time series visualization

Moreover, for visual assessment of any abnormality and presence of correlation between the recorded data on holidays and weekends, the time series of recorded data with corresponding mean and standard deviations are plotted. For this purpose, we investigated "grid_5161" (Figure 4.3) which is near the city center of Milan, and "grid_7524" (Figure 4.4) which is located in a suburban area in the northwest of Milan city. The weekends are highlighted with blue

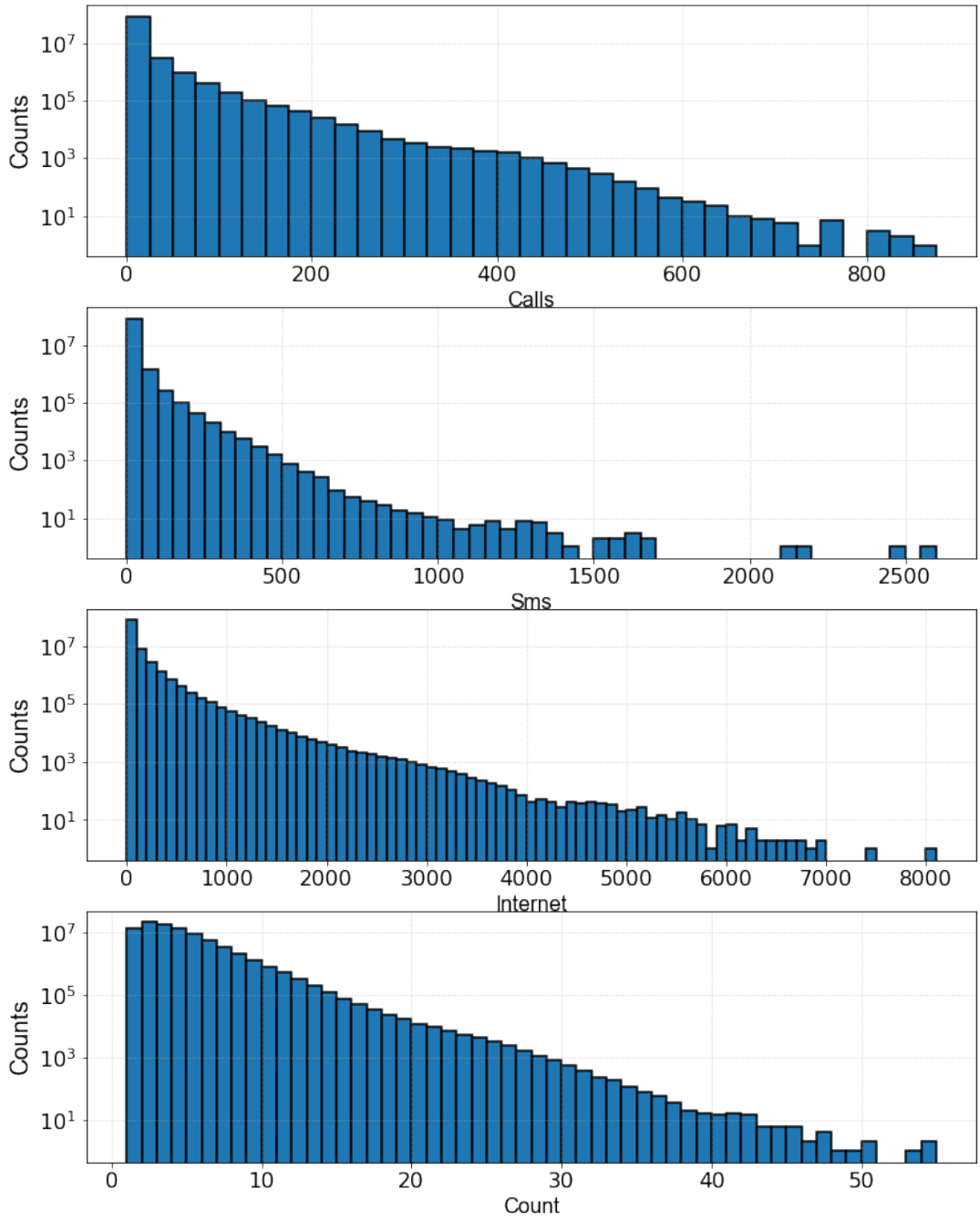


Figure 4.2: Skewness and kurtosis of features

and the holidays with pink. In the "grid_5161", the overall average traffic is higher compared to "grid_7524". For instance, the mean traffic of count is 14.626 for "grid_5161" whereas the "grid_7524" shows less mean traffic of count equal to 4.043. Also, In "grid_5161", we observe increasing traffic during weekends.

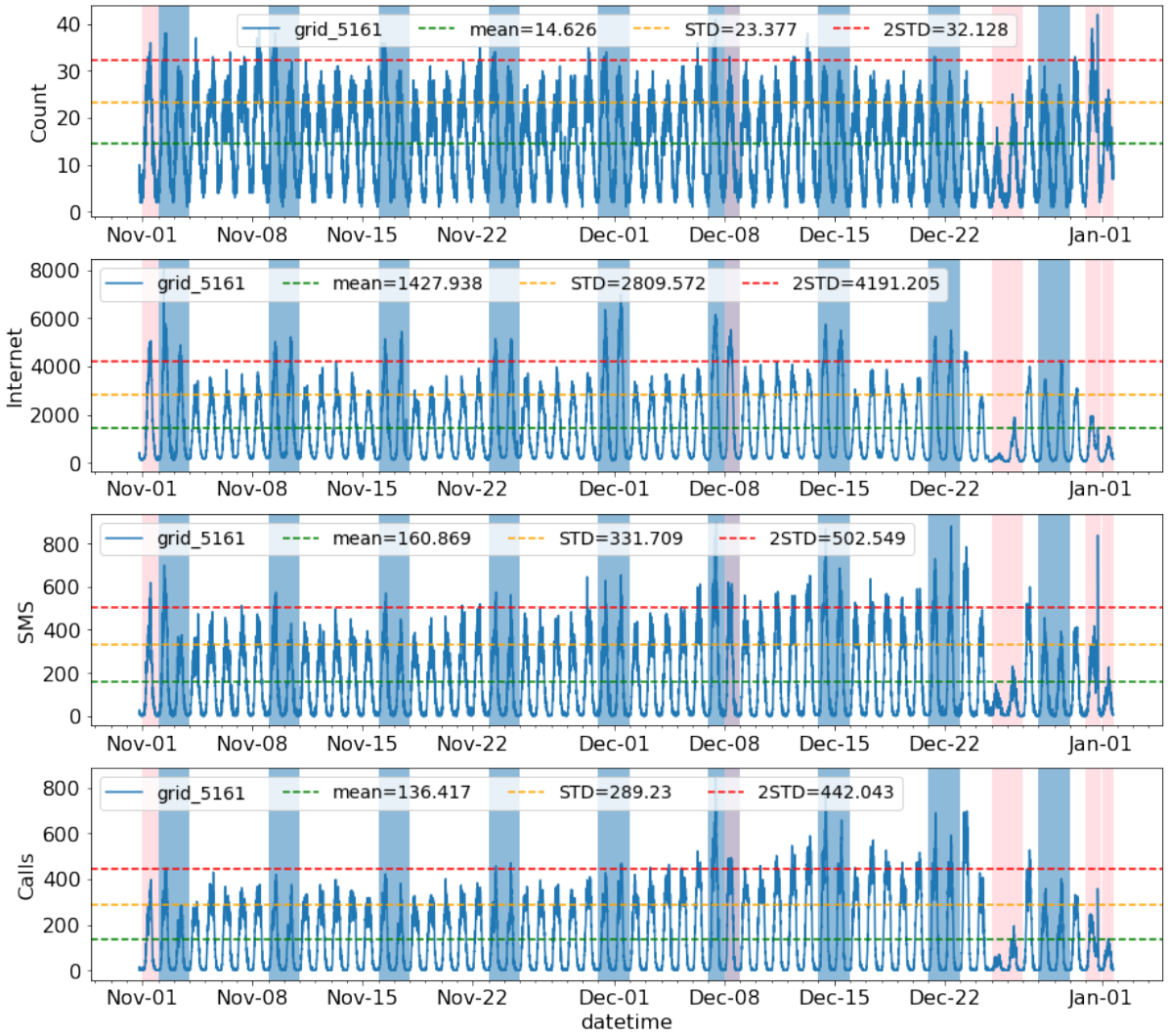


Figure 4.3: Time series visualization of 'grid_5161"

#### 4.4.2.5 Distribution analysis

Furthermore, the probability density function (PDF) of sms, count, calls, and internet for the "grid_5161" and "grid_7524" are illustrated in Figure 4.5. For the "grid_5161" we see a bi-modal distribution that has two peaks. Whereas, in the "grid_7524", since there is a sudden

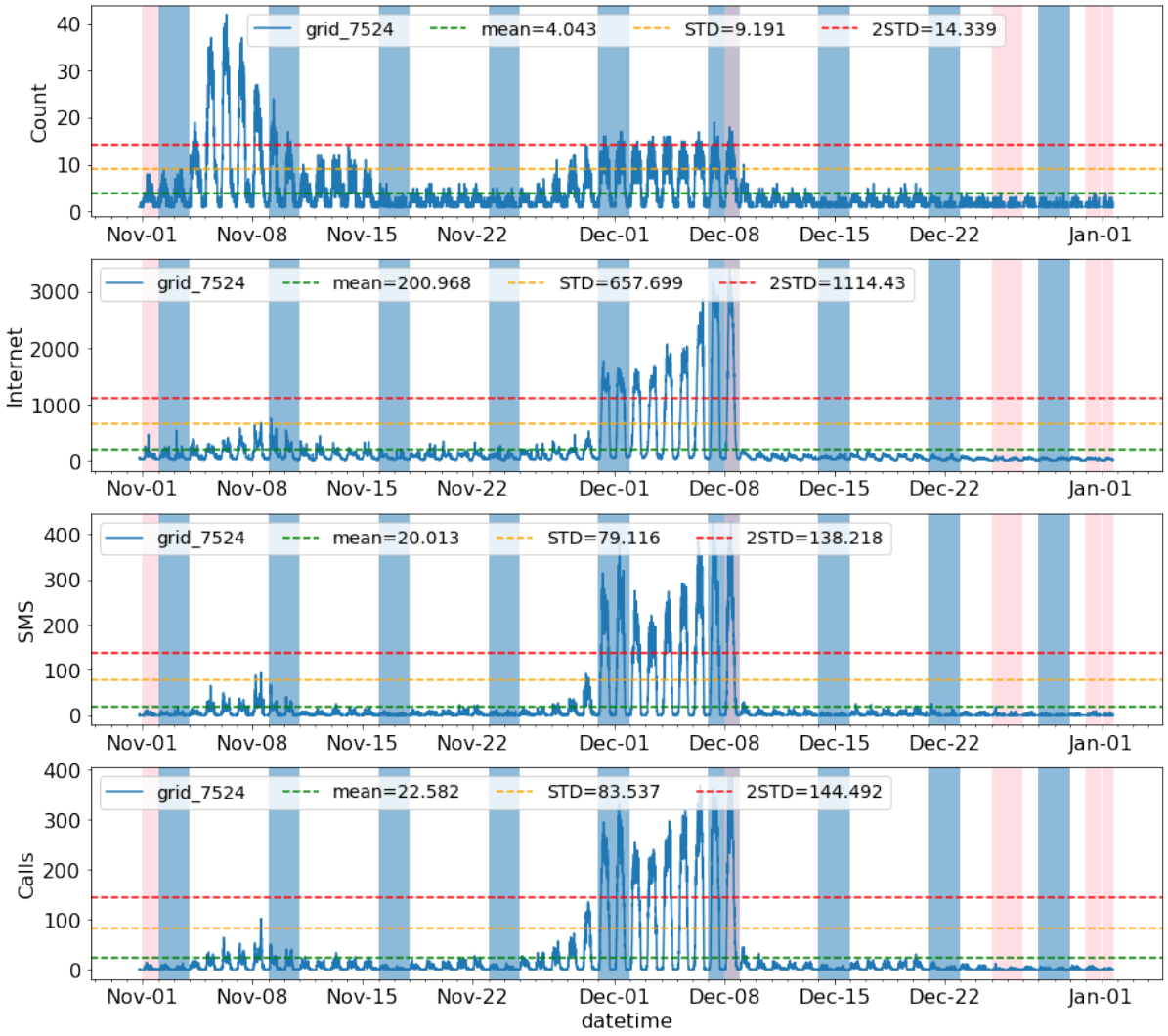


Figure 4.4: Time series visualization of 'grid_7524"

increase in particular days, we do not observe bimodality.

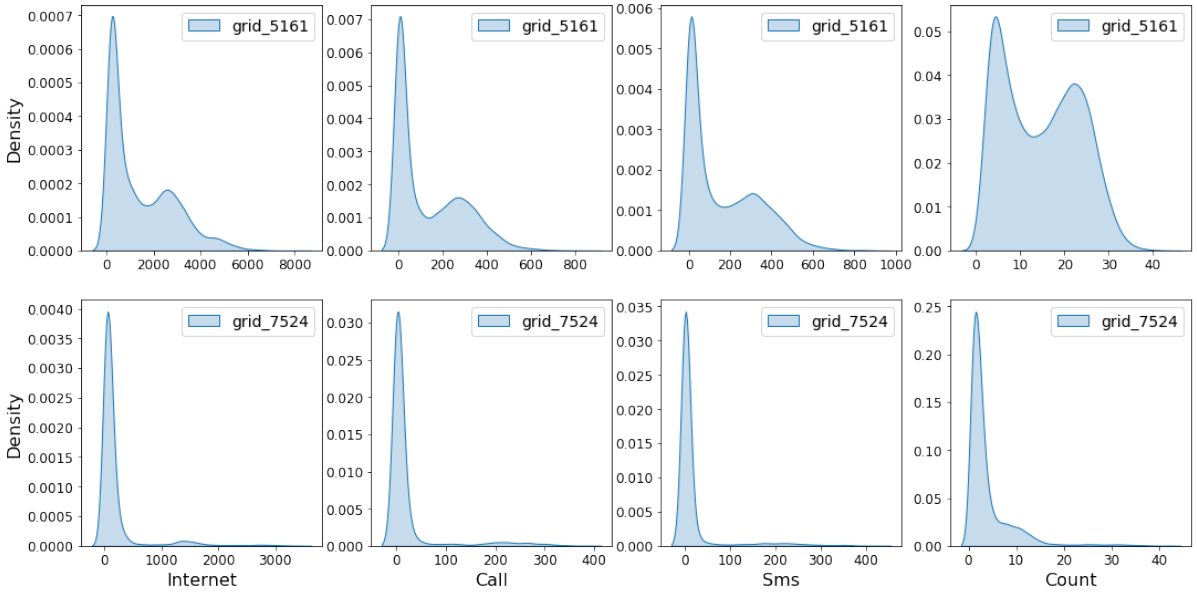


Figure 4.5: PDF of 'grid_5161" and "grid_7524"

For further analysis, we created a data frame for the average of internet, sms, and calls for all the grid IDs. Also, the histograms of the features, based on their average, are shown which are right-skewed (Figure 4.6). For each cell, the spatial-temporal average traffic of each feature is illustrated in Figure 4.7. The areas with a higher average are indicated by red. So, we can see the hot-spots for traffics in Milan city center with a significant reduction in the activity as we move away from the city center.

#### 4.4.2.6 Temporal dataset

Since the datetime in string form is not useful, the "datetime" values are converted to seconds. However, the data has an obvious daily and weekly periodicity. To deal with this issue, we adopted the time-frequency representation as follows.

$$f^s(\xi) = \sin \frac{2\pi t}{P}, \tag{4.1}$$

$$f^c(\xi) = \cos \frac{2\pi t}{P}, \tag{4.2}$$

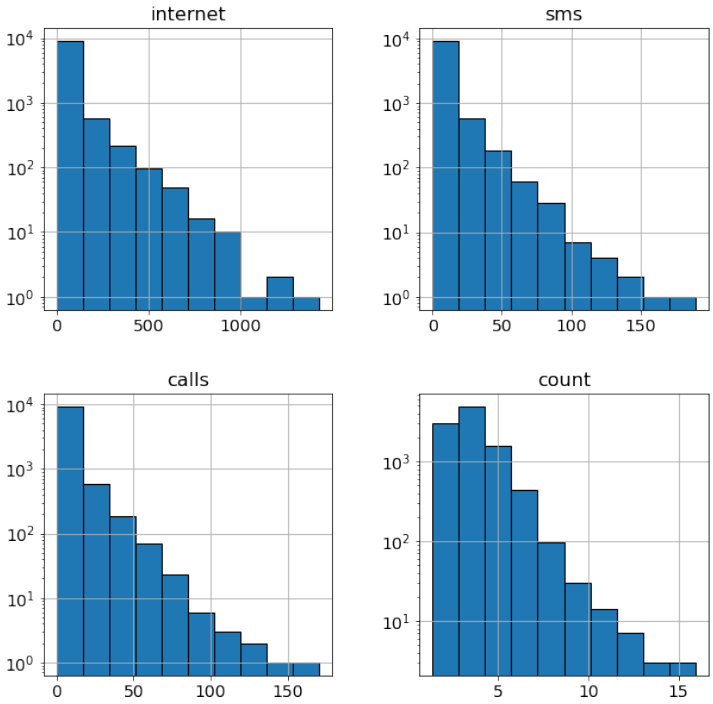


Figure 4.6: Average traffic of each grid

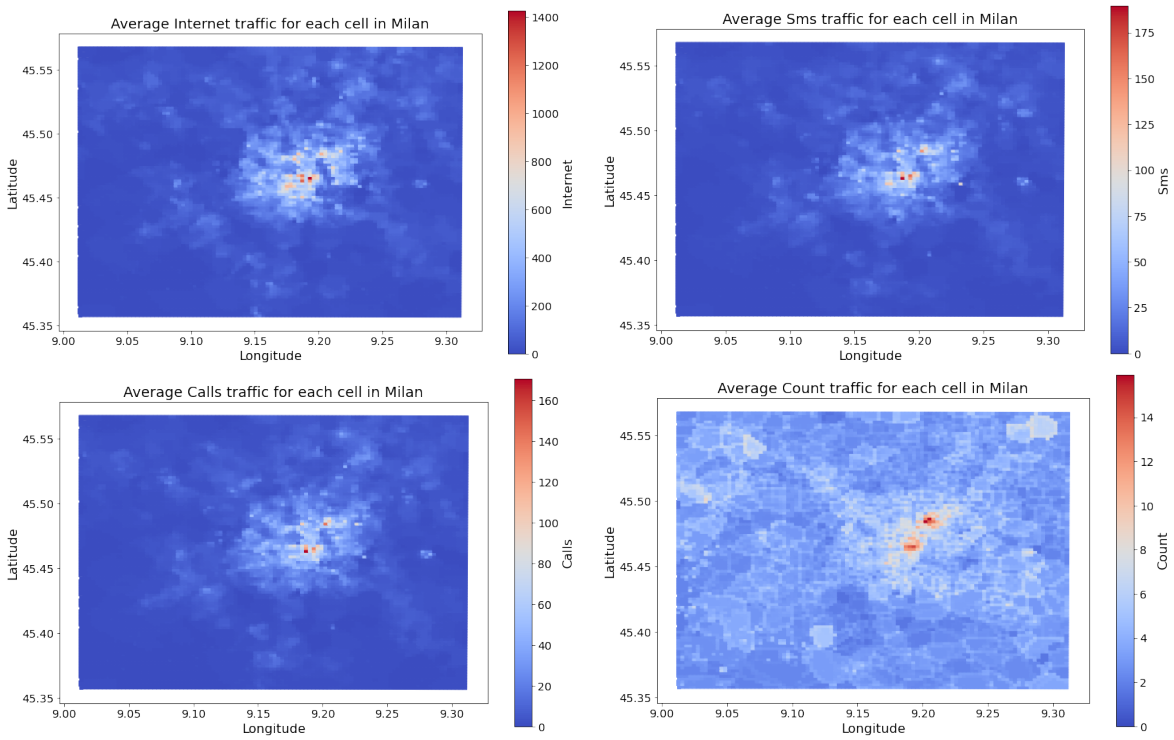


Figure 4.7: The spatial average traffic of each feature

where $t$ is time in second and $P$ is the length of the analyzed cycle (daily = 86400s, weekly = 604800s). In this temporal dataset, “internet”, “count”, “sms”, “calls”, “Day sin”, “Day cos”, “week sin”, and “week cos” are used as input to predict the cellular traffic of each grid for the next 24 hours of internet, count, sms, calls and frequency features based on the previous 24 hours.

#### 4.4.2.7 Spatiotemporal dataset

In the spatiotemporal dataset, the observation data for a specific time is like a frame with $100 \times 100$ pixels. In fact, the shape of the spatiotemporal dataset is like (1487,100, 100, 4) in which 1487 shows the time step (hour), $100 \times 100$ shows the longitude and latitude of grids and 4 illustrates the channels including the internet, count, sms, and call as model input. Also, Min-Max scaling was applied to the dataset for re-scaling the range of features between the range in [0, 1]. In Figure 4.8, the x_grid and y_grid which contains $100 \times 100$ grids (in total 10000 grids), and the time steps from the 1 to 1487 hours, are illustrated. Moreover, on the right side, the frame of the last time step is shown where each cube contains the records of internet, count, sms, and call. Also, Figure 4.9 illustrates internet usage of time steps 1 to 4. Higher internet consumption is shown in lighter colors while the low internet usage is depicted in darker colors.

### 4.4.3 Predictive Models

#### 4.4.3.1 Temporal models

In this work we explored the performance of five predictive models to forecast the telecommunication traffic for the next 24 hours based on the previous 24-hour data. For this purpose, several neural network-based models, as well as a baseline model, are utilized. The proposed approaches including Baseline model, FCSN, 1D-CNN, SS-LSTM, and AR-LSTM are elaborated as follows.

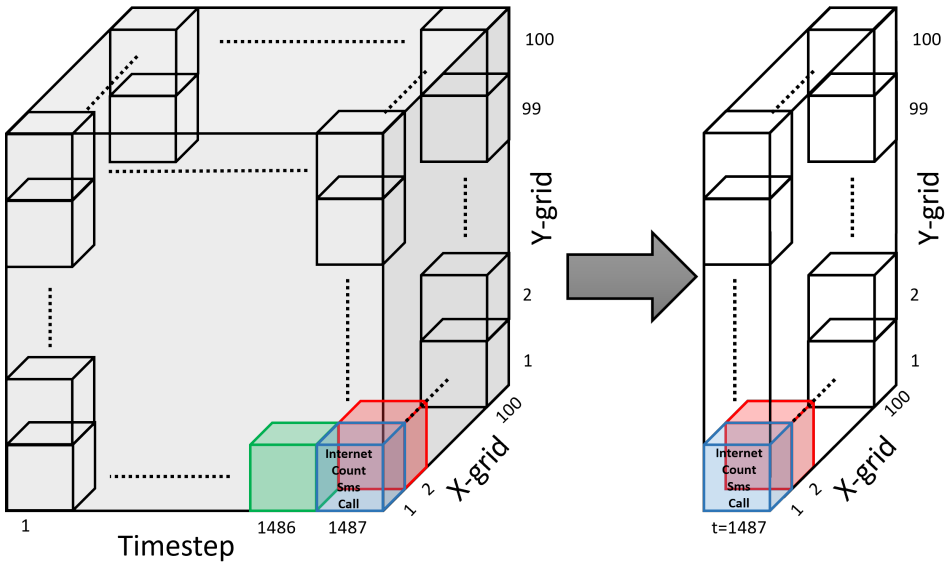


Figure 4.8: Schematic of spatiotemporal data

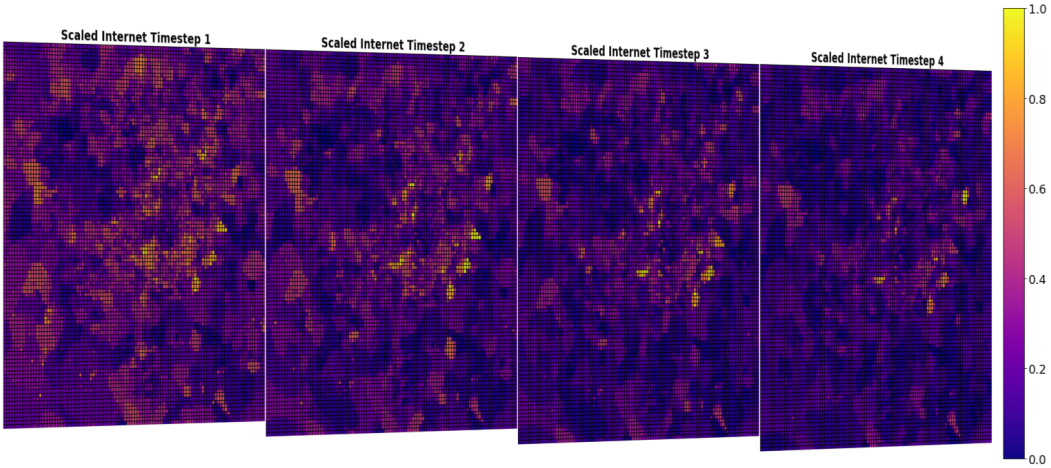


Figure 4.9: Frames of internet usage form time step 1 to 4

#### 4.4.3.1.1 Temporal baseline model

It is beneficial to establish a baseline to compare the performance of machine learning models. For the temporal baseline, it is assumed that the predicted 24 hours would have a similar pattern as the previous 24 hours. The adopted baseline model is shown in Figure 4.10.

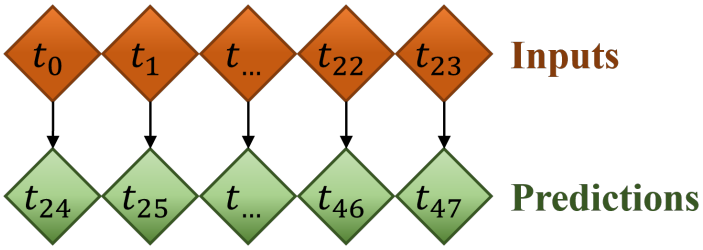


Figure 4.10: Block diagram of Baseline Structure.

#### 4.4.3.1.2 FCSN

A fully connected neural network is made up of a sequence of fully linked layers, where every neuron in one layer is connected to every neuron in other layers. The main benefit of the fully connected networks is that they are "structure agnostic," which means no particular hypotheses about the inputs are required [106]. We proposed the FCSN which stacked two dense layers with 512 and 192 nodes between the input and the output as shown in Figure 4.11. Also, rectified linear unit activation (RELU) was used to learn complex patterns in the data. Also, the graph in Figure 4.12 illustrates the connections in the neural networks. Each node in this diagram is labeled with the shape of its input and output matrices.

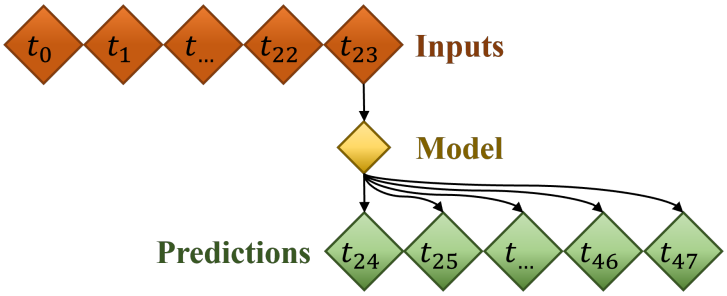


Figure 4.11: Block diagram of FCSN Structure.

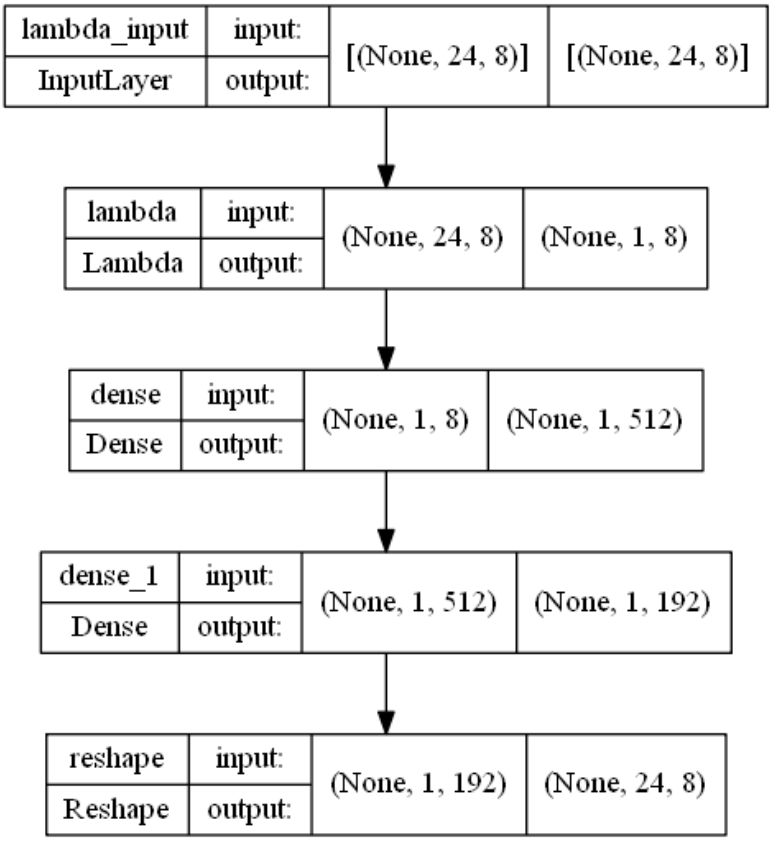


Figure 4.12: Graph visualization of FCSN model.

#### 4.4.3.1.3 1D-CNN

A one-dimensional convolution network produces an output tensor by convolving the input with the convolution kernel over a single dimension. Convolutional models create predictions based on a fixed-width history, which may yield better results than dense models since they can observe the changes over time. In this study, a 1D-CNN with a kernel size of 6 and a RELU activation function was developed (Figure 4.13) to predict the telecommunication traffic of the next 24 hours. I also used Graphviz to depict the connections between the neural networks. Each node in this diagram is labeled with the shape of its input and output matrices which is shown in the Figure 4.14.

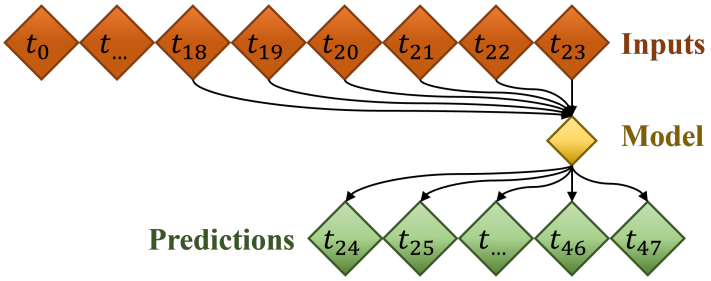


Figure 4.13: Block diagram of 1D-CNN Structure.

#### 4.4.3.1.4 SS-LSTM

The RNN architecture is well suited for processing a sequence of data, such as time series, since it makes use of memory in cells and takes data history into account (Figure 4.15). LSTM networks are a specific type of RNN which are developed to prevent the vanishing gradient problem. By utilizing the LSTM cells long-term dependencies can be learnt from data [107]. The structure of the LSTM units allows for learning the long-term dependencies. Unlike ordinary neurons, LSTM contains gates that control the learning process. Through the use of structures known as gates, each LSTM cell may store or forget information about previous network states. Due to its architecture, each memory cell's output is impacted by the sequence of previous states which makes LSTM appropriate for processing time series with long-time dependencies. Also, LSTM has been shown to be effective in language translation and speech recognition by utilizing a long history of inputs. The model will gather an internal state for 24

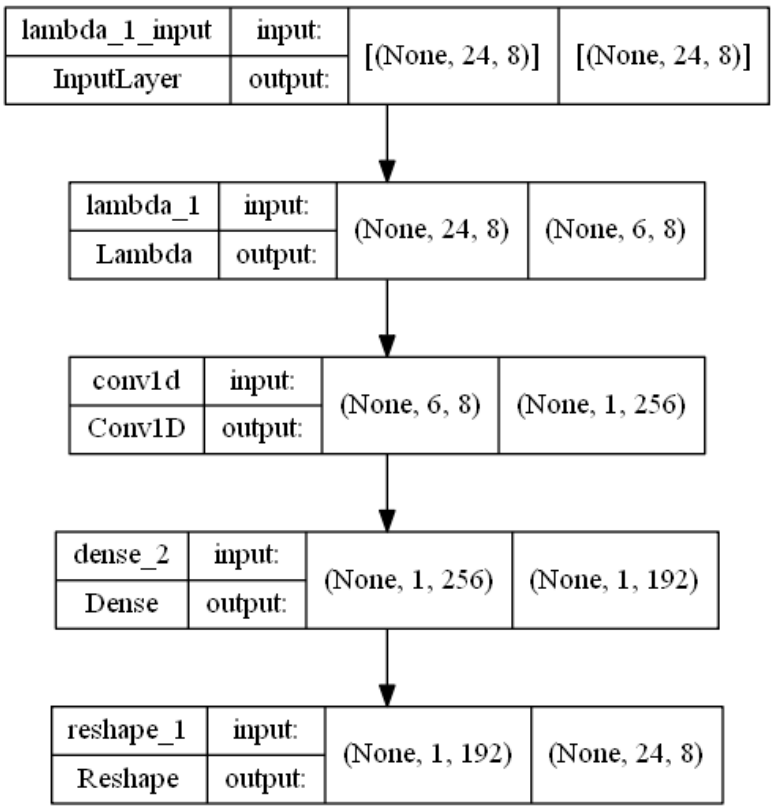


Figure 4.14: Graph visualization of 1D-CNN model.

hours before providing a single forecast for the following 24 hours.

In our proposed single-shot learning approach, LSTM only has to generate an output at the last time step. In this study, an LSTM layer with 32 internal units processes the sequence of inputs to predict telecommunication traffic over the next 24 hours. In particular, data from the previous 24 hours were used as input for a single-shot prediction of the next 24 hours sequence. The graph in Figure 4.16 shows the connections in the LSTM network. Each node in this diagram is labeled with the shape of its input and output matrices.

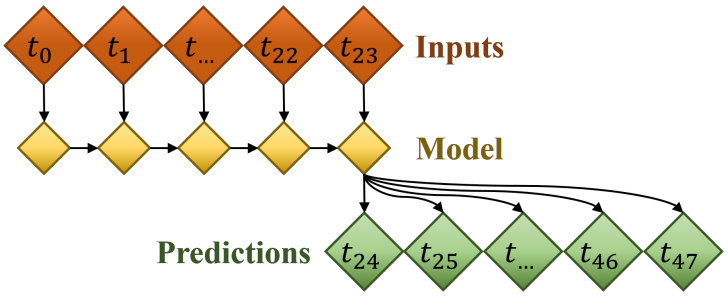


Figure 4.15: Block diagram of SS-LSTM Structure.

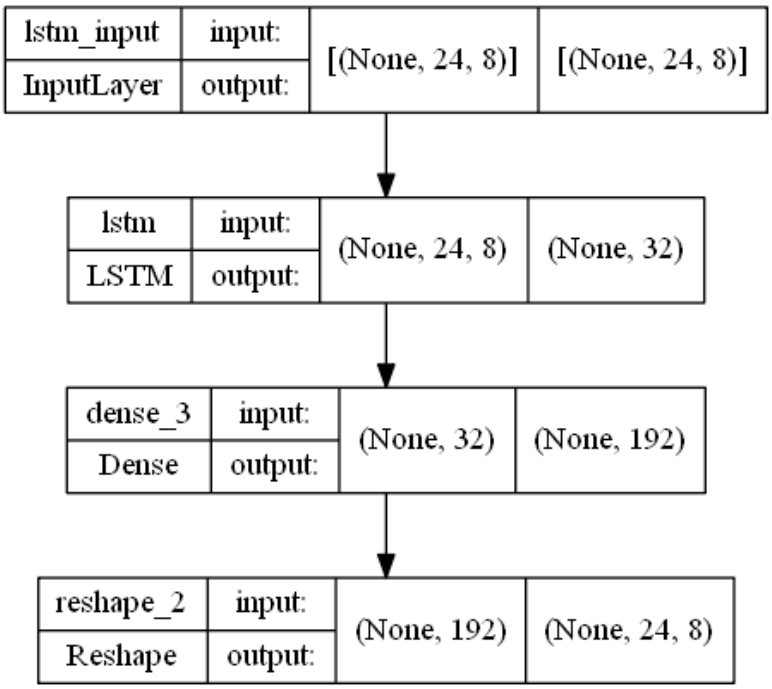


Figure 4.16: Graph visualization of SS-LSTM model.

#### 4.4.3.1.5 AR-LSTM

All of the models listed above, anticipate the whole output sequence in one single step. It may be beneficial for the model to breakdown this prediction into separate time. Like in the classic RNN, for generating sequences, each model's output may be fed back into itself at each phase, and predictions can be made based on the preceding ones (Figure 4.17). One advantage of this type of model is that it can be adjusted to generate output of varied lengths. In this work, we proposed a sequential model, by wrapping an LSTM cell layer in the lower level of the RNN layer to simplify the "warmup" method to predict telecommunication traffic over the next 24 hours. The warmup method returns a single time step prediction and the internal state of the LSTM. With the RNN's state, and an initial prediction we can continue iterating the model and feed the predictions at each time step as input to predict the next time steps. Moreover, the AR-LSTM model summary is illustrated in Figure 4.18. The model summary contains information on the layers and their order in the model, the output shape of each layer, the number of parameters in each layer, and the overall number of parameters in the model.

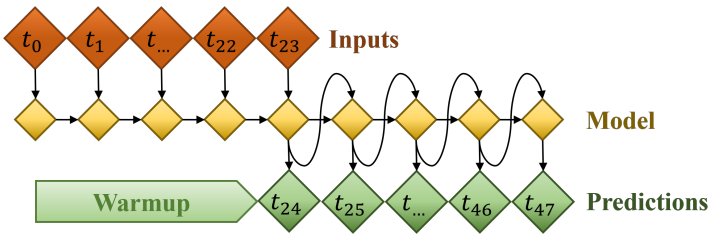


Figure 4.17: Block diagram of AR-LSTM Structure.

### 4.4.3.2 Spatiotemporal models

A 2D-ConvLSTM model is proposed for the a multivariate spatiotemporal analysis to forecast cellular traffic over the next 24 hours. Moreover, the spatiotemporal baseline is used as a benchmark to evaluate the effectiveness of the 2D-ConvLSTM model.

#### 4.4.3.2.1 Spatiotemporal baseline model

Before developing the spatiotemporal model, it is helpful to create a performance baseline for comparison. The baseline assumes that the anticipated 24 hours will follow a pattern similar to

```
Model: "AR-LSTM"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 lstm_cell_1 (LSTMCell)       multiple                  5248

 rnn (RNN)                    multiple                  5248

 dense_6 (Dense)              multiple                  264

=================================================================
Total params: 5,512
Trainable params: 5,512
Non-trainable params: 0
_________________________________________________________________
None
```

Figure 4.18: Summary of AR-LSTM model.

the preceding 24 hours.

#### 4.4.3.2.2 2D-ConvLSTM

In order to extract spatial and temporal dependencies of data simultaneously and incorporate them in the traffic prediction, a 2D-ConvLSTM network is proposed to analyze the multi-channel spatiotemporal data. The proposed 2D-ConvLSTM framework (Figure 4.19) consists of 4 layers of 2D Convolutional LSTM and one 3D convolutional layer. The framework of the proposed spatiotemporal model is illustrated in Figure 4.19. As for the input, the data with the shape of (24,100,100,4) are considered to conduct the multivariate analysis and incorporate correlation among variables, space, and time. Precisely, input data comprise the 24 hours records across all grids for the 4 channels including internet, count, sms, and call records. In the proposed model, input data with the aforementioned shape is fed into the 2D Convolutional LSTM layer which extracts features in 10 channels. In the next layer, the obtained 10 channels are shrunk via progressive channel sorting to 8, 5 and 3 channels in the next layers, respectively, for optimization and reducing the number of parameters. Extracted features in each layer are normalized for mean centering and variance scaling. This batch normalization reduces the risk of internal co-variate shift in later layers. The obtained 3 channels are then fed into a 3D convolutional layer in the last layer to predict the next cellular traffic frame for each individual

data type, $d$. In fact, $d \in \{internet, sms, call\}$ indicates specific kinds of cellular traffic that are forecasted individually with the proposed spatiotemporal 2D-ConvLSTM model. Moreover, the summary of the proposed model is shown in Figure 4.20.

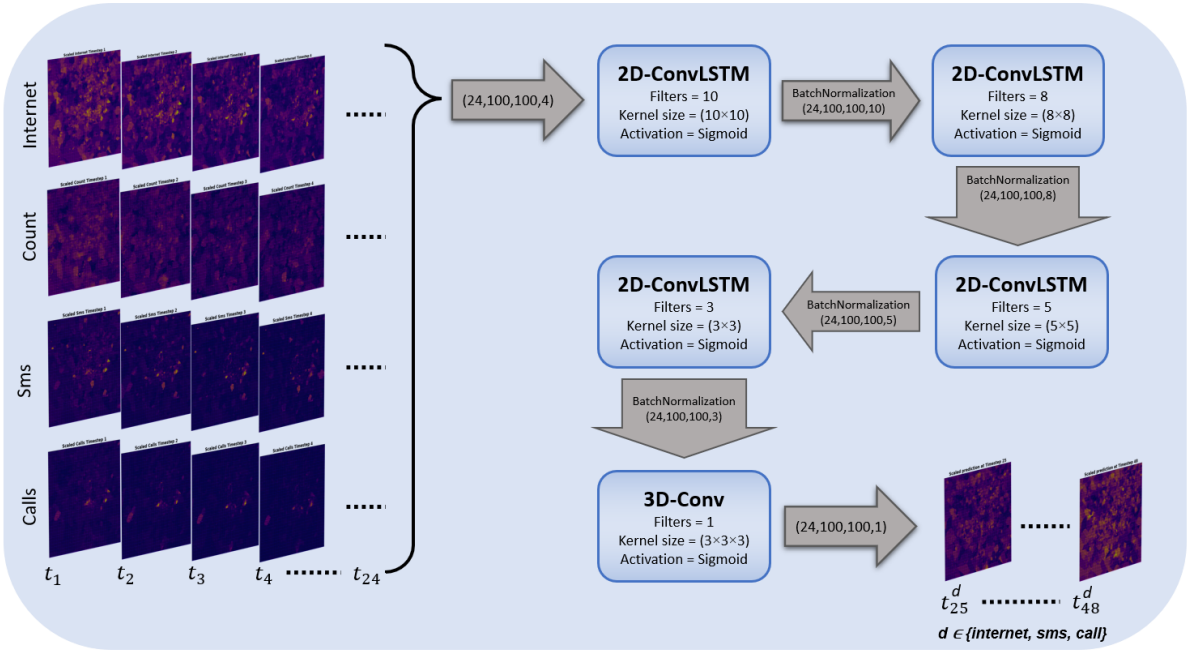


Figure 4.19: Framework of 2D-ConvLSTM model.

## 4.5 Results and Evaluation

### 4.5.1 Evaluation metrics

In order to assess the performance of the proposed predictive models compared to other works, in this study the temporal and spatiotemporal models were evaluated by utilizing MAE and RMSE.

#### 4.5.1.1 Mean absolute error (MAE)

The average size of the errors in a series of predictions is measured by MAE. The absolute differences between model prediction and the actual observation are averaged over the test

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 conv_lstm2d_4 (ConvLSTM2D)  (None, 24, 100, 100, 10)  56040

 batch_normalization_4 (Batc  (None, 24, 100, 100, 10)  40
 hNormalization)

 conv_lstm2d_5 (ConvLSTM2D)  (None, 24, 100, 100, 8)   36896

 batch_normalization_5 (Batc  (None, 24, 100, 100, 8)  32
 hNormalization)

 conv_lstm2d_6 (ConvLSTM2D)  (None, 24, 100, 100, 5)   6520

 batch_normalization_6 (Batc  (None, 24, 100, 100, 5)  20
 hNormalization)

 conv_lstm2d_7 (ConvLSTM2D)  (None, 24, 100, 100, 3)   876

 batch_normalization_7 (Batc  (None, 24, 100, 100, 3)  12
 hNormalization)

 conv3d_1 (Conv3D)           (None, 24, 100, 100, 1)   82

=================================================================
Total params: 100,518
Trainable params: 100,466
Non-trainable params: 52
_________________________________________________________________
```

Figure 4.20: Summary of 2D-ConvLSTM model.

sample by MAE.

$$\textbf{MAE} = (\frac{1}{n}) \sum_{i=1}^{n} |y_i - \hat{y}_i| \ , \tag{4.3}$$

where $n$ shows the total number of data points, $y_i$ indicates the true output value, and $\hat{y}_i$ represents the predicted value for $i^{th}$ data point.

#### 4.5.1.2 Root mean squared error (RMSE)

The RMSE is a quadratic scoring rule that determines the average magnitude of errors. RMSE is the square root of the average of squared differences between predicted and observed values.

$$\textbf{RMSE} = \sqrt{(\frac{1}{n}) \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \, , \tag{4.4}$$

where $n$ represents the total number of data points, $y_i$ indicates the true output value, and $\hat{y}_i$ represents the predicted value for $i^{th}$ data point. RMSE is more sensitive to the data with a larger difference between actual and predicted values. This is because the error is squared before the average is reduced with the square root. While MAE is more robust to outliers, RMSE is more sensitive to outliers compared to the MAE. In fact, data with higher errors would skew the RMSE.

### 4.5.2 Temporal cellular traffic prediction

Each grid contains 1487 records. For each grid, We allocated 70% of the dataset, 1040 records, to the training set, 20%, 298 records, to the validation set, and the remaining 10% of the dataset, corresponding to 149 records, were devoted to the test set. For instance, the splitting of the dataset for normalized internet in "grid_5161" is shown in Figure 4.21. Also, the data was preprocessed and normalized to zero mean and unit variance before feeding to the neural networks. The models were trained for 50 epochs using Adam optimizer and mean squared error (MSE) as the loss function. Moreover, hyperparameters were tuned to control the learning process toward an optimal prediction.

- **Temporal baseline:** The hyperparameters of the temporal baseline model are shown in

Table 4.3. The lowest validation loss was achieved by utilizing MSE as the loss function, Adam optimizer, and a learning rate of 0.001. Also, the temporal baseline is structured based on utilizing time step 1 to predict the first observed point, time step 2 for predicting the 2nd observed point and so on, until using time step 24 to predict the traffic on the 24$^{th}$ observed point.

Table 4.3: Temporal baseline hyperparameters.

| Criteria / Model | Loss | | Optimizer | | | Learning rate | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | MSE | Adam | Adamax | SGD | 0.1 | 0.01 | 0.001 |
| **Temporal baseline** | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ |
| **Model structure** | • Considering time step 23 to predict the next 24 hours.<br>• **Using the time step 1 to predict the first observed point, time step 2 for predicting the 2nd observed point and so on, until using time step 24 to predict the traffic of 24$^{th}$ observed point.** | | | | | | | |

- **FCSN:** The hyperparameters of the FCSN model are illustrated in Table 4.4. The lowest validation loss was obtained by using MSE as the loss function, Adam optimizer, and a learning rate of 0.001. The best prediction results were obtained by utilizing RELU activation and 512 neurons in the hidden layer.

Table 4.4: FCSN hyperparameters.

| Criteria / Model | Loss | | Optimizer | | | Learning rate | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | MSE | Adam | Adamax | SGD | 0.1 | 0.01 | 0.001 |
| **FCSN** | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ |
| **Model structure** | • Considering **RELU**, leaky RELU, and Tanh activation functions.<br>• Considering 256, **512**, and 1024 units for hidden layer. | | | | | | | |

- **1D-CNN:** Table 4.5 shows the hyperparameter used in the 1D-CNN model. By using MSE as the loss function, Adam optimizer, and the learning rate of 0.001, the lowest validation loss was observed. 1D-CNN achieved the best results by utilizing the kernel size of 6, RELU activation function, and 256 filters in the Conv1D layer.

- **SS-LSTM:** The hyperparameter tuning of the SS-LSTM model is illustrated in Table 4.6. The validation loss had the lowest amount when using MSE as a loss function, Adam

Table 4.5: 1D-CNN hyperparameters.

| Criteria / Model | Loss | | Optimizer | | | Learning rate | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | MSE | Adam | Adamax | SGD | 0.1 | 0.01 | 0.001 |
| **1D-CNN** | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ |
| **Model structure** | • Considering kernel size of 4, 5, **6**.<br>• Considering **RELU**, Tanh, and sigmoid activation functions.<br>• Considering 128, **256**, and 512 filters in Conv1D layer. | | | | | | | |

optimizer, and a learning rate of 0.001. Moreover, the best results were obtained by using 32 LSTM units and no dropout.

Table 4.6: SS-LSTM hyperparameters.

| Criteria / Model | Loss | | Optimizer | | | Learning rate | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | MSE | Adam | Adamax | SGD | 0.1 | 0.01 | 0.001 |
| **SS-LSTM** | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ |
| **Model structure** | • Considering **32**, 64, 128, 256 LSTM units.<br>• Considering **0**, 0.2 and 1 for the dropout. | | | | | | | |

- **AR-LSTM:** Table 4.7 illustrates that the lowest validation loss in the AR-LSTM model is obtained by using MSE as a loss function, Adam optimizer, learning rate of 0.001 and 32 LSTM units.

Table 4.7: AR-LSTM hyperparameters.

| Criteria / Model | Loss | | Optimizer | | | Learning rate | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | MSE | Adam | Adamax | SGD | 0.1 | 0.01 | 0.001 |
| **SS-LSTM** | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ |
| **Model structure** | • Considering **32**, 64, 128, 256 LSTM units. | | | | | | | |

The predictions by the baseline, FCSN, 1D-CNN, SS-LSTM, and AR-LSTM models in the "grid_5161" are illustrated in Figure 4.22, Figure 4.23, Figure 4.24, Figure 4.25, and Figure 4.26 respectively. Moreover, the overall performance of the prediction scheme on the test set is evaluated.

The averaged MAE is calculated for the features within all 10000 grids which is shown in Table 4.8. Although FCSN and 1D-CNN have comparable performance, 1D-CNN is a smaller

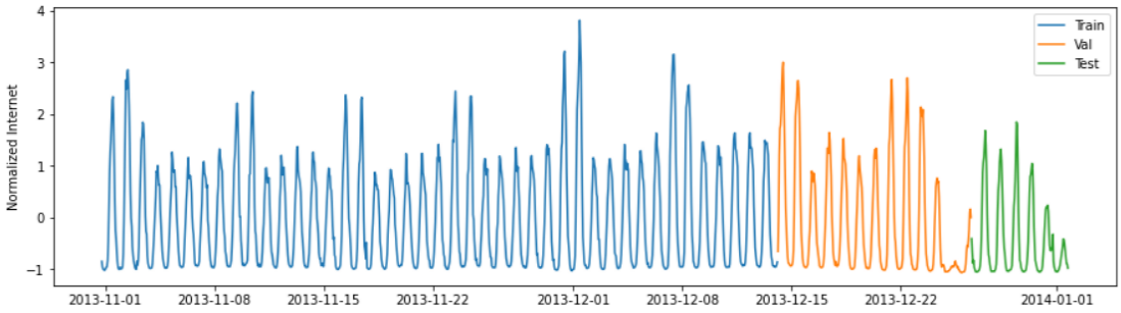


Figure 4.21: Illustration of train, validation, and test set split for the normalized internet traffic of "grid_5161"

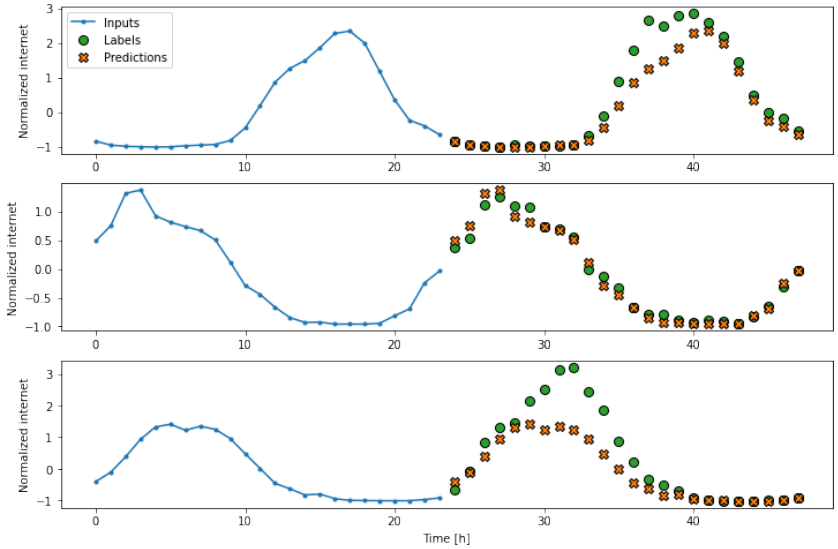


Figure 4.22: Baseline model prediction on normalized internet traffic of "grid_5161".

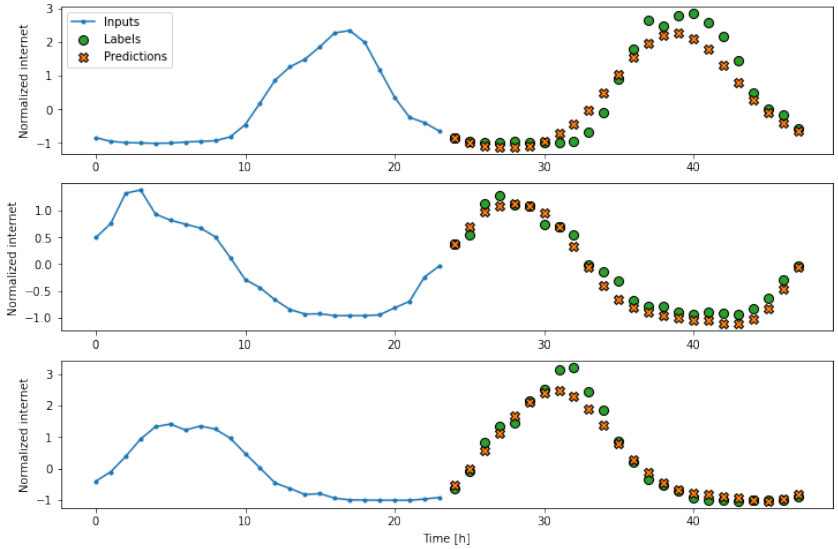


Figure 4.23: FCSN model prediction on normalized internet traffic of "grid_5161".

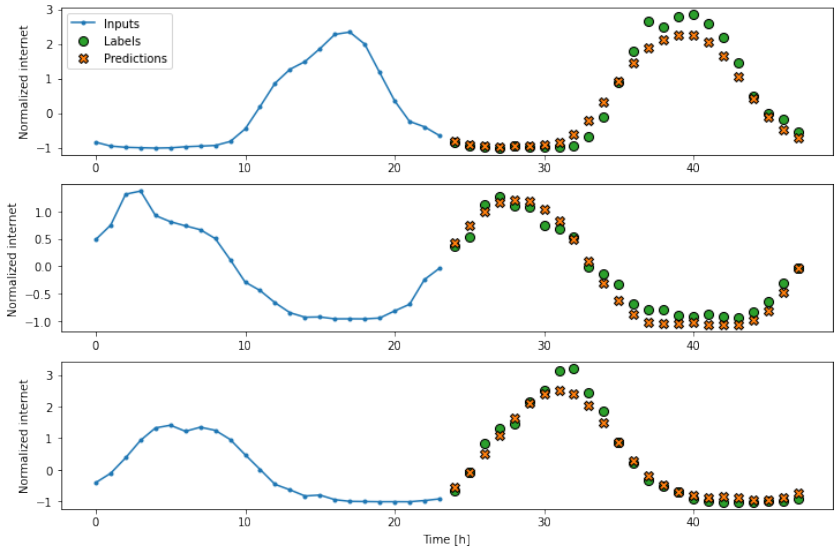


Figure 4.24: 1D-CNN model prediction on normalized internet traffic of "grid_5161".

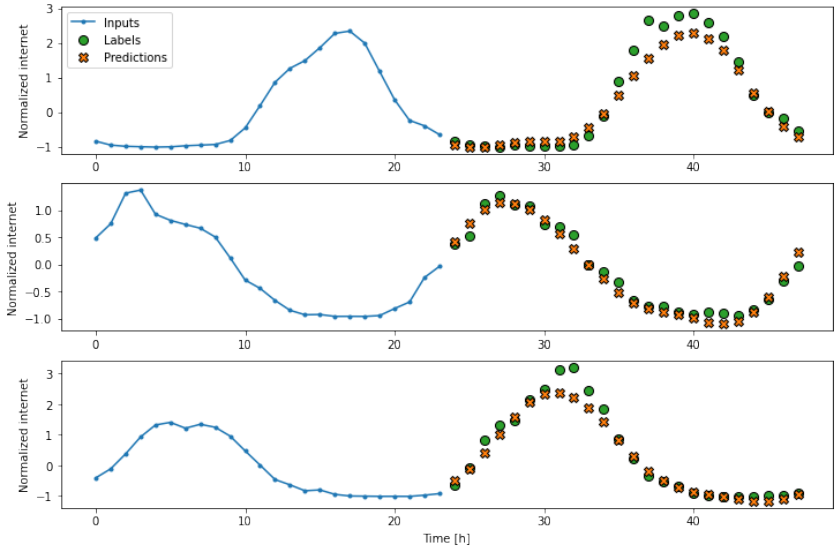


Figure 4.25: SS-LSTM model prediction on normalized internet traffic of "grid_5161".

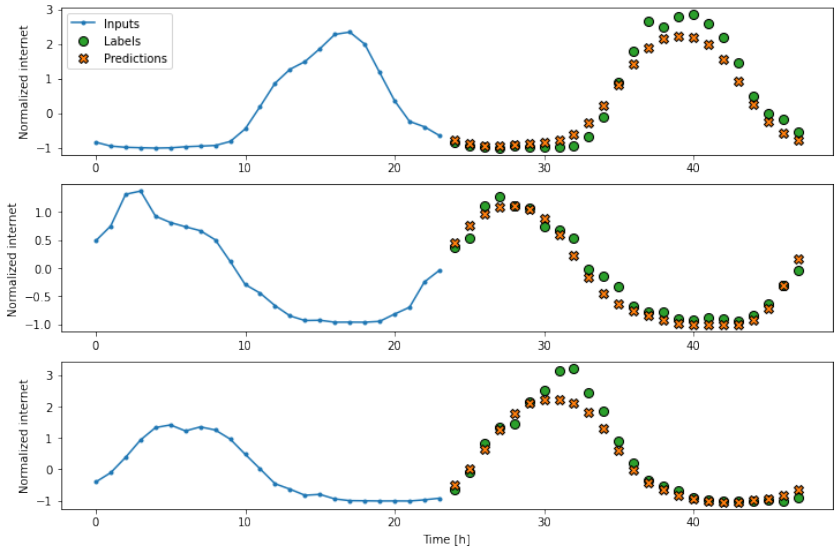


Figure 4.26: AR-LSTM model prediction on normalized internet traffic of "grid_5161".

network with less number of parameters. Also, Figure 4.27 illustrates the MAE distribution for each model over all grids.

Table 4.8: Averaged MAE for all features across all grids

| Model | Averaged MAE |
|---|---|
| **Baseline Model** | 0.39 |
| **FCSN** | **0.29** |
| **1D-CNN** | **0.29** |
| **SS-LSTM** | 0.32 |
| **AR-LSTM** | 0.32 |

Moreover, the prediction performance of temporal models including temporal baseline, FCSN, 1D-CNN, and SS-LSTM are investigated. The AR-LSTM model is not used for the prediction of specific types of cellular traffic. The AR-LSTM model fails in mobile traffic forecasting when the number of features is decreased as the model outputs are fed back at each phase to be used as input for the predictions of the next time steps. The cellular traffic forecasting of temporal models over specific types of cellular traffic including internet, sms, and calls are summarized in the following Table 4.9. According to the Table 4.9, for predicting the internet traffic, 1D-CNN outperforms other models with the smallest MAE and RMSE. However, for sms prediction, all three models have comparable performance. In terms of call

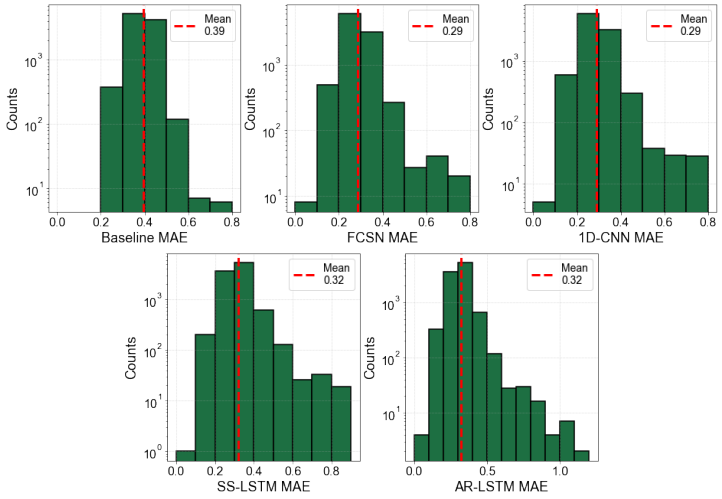


Figure 4.27: MAE distribution of the proposed models over all grids.

traffic prediction, the performance of the SS-LSTM model is slightly better than other temporal models. However, the 1D-CNN model has smaller model size with less complexity and it is more adequate to be used for edge computing deployment.

Table 4.9: Cellular traffic predictions of temporal models

| Traffic | Model | MAE | RMSE |
|---|---|---|---|
| **Internet** | Temporal baseline | 153.72 | 205.95 |
| | FCSN | 117.39 | 152.66 |
| | 1D-CNN | **113.54** | **147.43** |
| | SS-LSTM | 124.32 | 160.93 |
| **Sms** | Temporal baseline | 30.07 | 45.56 |
| | FCSN | 17.96 | **32.30** |
| | 1D-CNN | 17.60 | 32.62 |
| | SS-LSTM | **17.10** | 32.66 |
| **Call** | Temporal baseline | 27.02 | 36.62 |
| | FCSN | 14.92 | 22.42 |
| | 1D-CNN | 14.77 | 22.26 |
| | SS-LSTM | **13.08** | **21.25** |

Furthermore, the spatial distribution of the averaged MAE of the FCSN predictions is demonstrated in Figure 4.28. In this spatial distribution, high MAEs are indicated in red and low MAEs are depicted in blue. Hence, darker blue colors show fewer errors.

Finally, in order to have comprehensive assessment of the proposed forecasting models, the execution time of the models are illustrated in Table 4.10. As we expected, the baseline model

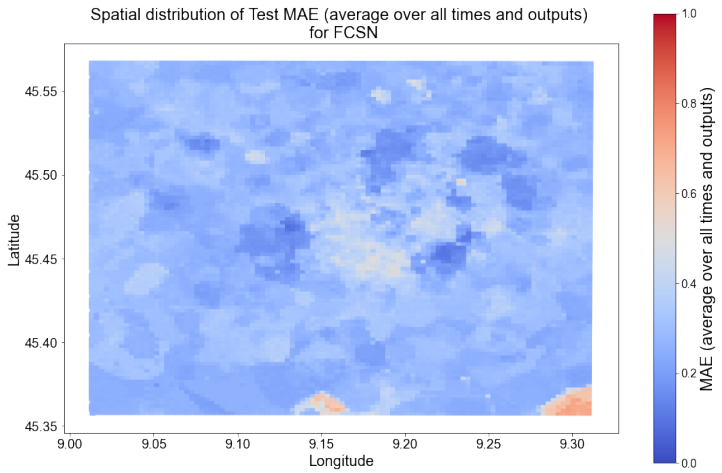


Figure 4.28: Spatial distribution of MAE for FCSN model predictions.

is the fastest with 0.31 seconds run-time, and the most complex model, AR-LSTM, has the longest execution time of 24.31 seconds. Table 4.10 confirms that FCSN and 1D-CNN have significant difference in execution time compared to LSTM approaches.

Table 4.10: Execution time of the proposed predictive models

| Model | Execution time (seconds) |
|---|---|
| **Baseline Model** | 0.31 |
| **FCSN** | 6.98 |
| **1D-CNN** | 6.37 |
| **SS-LSTM** | 21.38 |
| **AR-LSTM** | 24.31 |

### 4.5.3 Spatiotemporal cellular traffic prediction

For spatiotemporal cellular traffic prediction, we devoted 70% of the dataset (1040 CDRs) for the training set, 20% (298 records) for the validation set, and 10% (149 records) for the test set. Moreover, Min-Max scaling was applied to the dataset for rescaling the range of inputs between the range in [0, 1]. The 2D-ConvLSTM model was trained by considering 500 epochs with early stopping that monitors the validation loss. The early stopping interrupts the training process when the error on the validation set has not decreased for four epochs. Also, Adam optimizer and MAE as the loss function were used during the training process.

Furthermore, to control the learning process, the hyperparameters leveraged for the spatiotemporal baseline and the 2D-ConvLSTM model are presented below.

- **Spatiotemporal baseline:** Table 4.11 displays the hyperparameters of the spatiotemporal baseline model. With MAE as the loss function, Adam as the optimizer, and a learning rate of 0.001, the lowest validation loss was obtained. Furthermore, the spatiotemporal baseline is organized so that time step 1 is used to forecast the first observed point, time step 2 is used to predict the second observed point, and so on, until time step 24 is used to predict traffic on the $24^{th}$ observed point.

Table 4.11: Spatiotemporal baseline hyperparameters.

| **Criteria** / **Model** | **Loss** | | **Optimizer** | | | **Learning rate** | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | MSE | Adam | Adamax | SGD | 0.1 | 0.01 | 0.001 |
| **Spatiotemporal baseline** | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ |
| **Model structure** | • **Using the time step 1 to predict the first observed point, time step 2 for predicting the 2nd observed point and so on, until using time step 24 to predict the traffic of $24^{th}$ observed point.** | | | | | | | |

- **2D-ConvLSTM model:** The hyperparameter tuning process for the 2D-ConvLSTM model is depicted in Table 4.12. When utilizing MAE as a loss function, Adam optimizer, and a learning rate of 0.001, the validation loss is the smallest. The best results are seen by using sigmoid as an activation function, assuming 5 hidden layers. Moreover, kernel sizes including $(10 \times 10)$, $(8 \times 8)$, $(5 \times 5)$, and $(3 \times 3)$ for the aforementioned layers along with filter sizes of 10, 8, 5, and 3 lead to achieving the best results.

The predictions of the internet, sms, and call by the 2D-ConvLSTM model across all grids for the time step 61 are depicted in Figure 4.29, Figure 4.30, and Figure 4.31 respectively.

The cellular traffic performance of internet, sms, and calls for spatiotemporal baseline and 2D-ConvLSTM are presented in Table 4.13. Also, the execution time of the proposed models over all types of cellular traffic including internet, sms, and call are illustrated in Table 4.14.

Table 4.12: 2D-ConvLSTM hyperparameters.

| Criteria / Model | Loss | | Optimizer | | | Learning rate | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | MSE | Adam | Adamax | SGD | 0.1 | 0.01 | 0.001 |
| **2D-ConvLSTM** | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ |
| **Model structure** | • Considering RELU, Tanh, and **sigmoid** activation functions.<br>• Considering various numbers of hidden layers starting from 1, 2, 3, 4, and **5**.<br>• Considering different kernel sizes including **(10 × 10)**, **(8 × 8)** , (6 × 6), **(5 × 5)**, **(3 × 3)**.<br>• Considering different filter sizes containing **10**, **8**, 6, **5**, and **3**. | | | | | | | |

Table 4.13: Various types of cellular traffic performance

| Traffic | Model | MAE | RMSE |
|---|---|---|---|
| **Internet** | Spatiotemporal baseline | 102.12 | 142.41 |
| | **2D-ConvLSTM** | **52.73** | **75.73** |
| **Sms** | Spatiotemporal baseline | 24.03 | 36.04 |
| | **2D-ConvLSTM** | **14.42** | **26.60** |
| **Call** | Spatiotemporal baseline | 15.23 | 22.06 |
| | **2D-ConvLSTM** | **8.98** | **15.02** |

Table 4.14: Execution time of proposed models on various types of traffic

| Traffic | Model | Execution time (seconds) |
|---|---|---|
| **Internet** | Spatiotemporal baseline | 10.15 |
| | 2D-ConvLSTM | 4419.16 |
| **Sms** | Spatiotemporal baseline | 6.89 |
| | 2D-ConvLSTM | 2821.77 |
| **Call** | Spatiotemporal baseline | 12.87 |
| | 2D-ConvLSTM | 3406.08 |

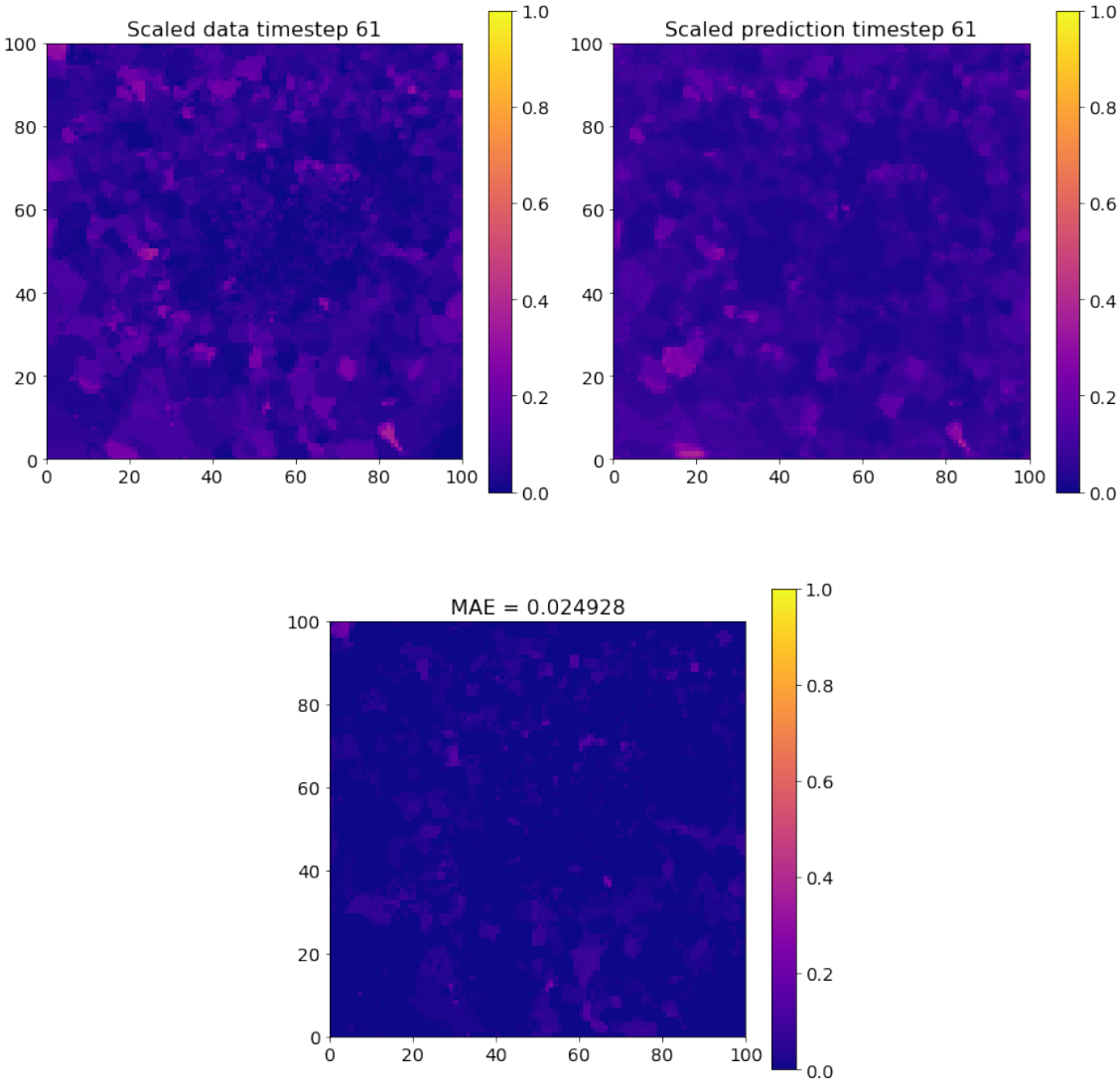


Figure 4.29: 2D-ConvLSTM internet prediction performance on time step 61 .

### 4.5.4 Spatiotemporal performance compared to the related works

To demonstrate the superiority of the proposed 2D-ConvLSTM model in forecasting cellular traffic, the performance of the 2D-ConvLSTM is compared with the STCNet which is proposed by Zhang *et al.* [103]. STCNet is a cross-domain analysis including social, BS, and POI data to capture external factors that affect cellular traffic generation. The comparison between the performance of our proposed model and STCNet on internet, call, and sms traffic prediction are illustrated in Table 4.15, Table 4.16 and Table 4.17 respectively.

Compared to the reference STCNet models [103], our proposed 2D-ConvLSTM shows the

Table 4.15: Comparison between models in predicting internet traffic

| Traffic | Model | MAE | RMSE |
|---|---|---|---|
| **Internet** | Spatiotemporal baseline | 102.12 | 142.41 |
| | **2D-ConvLSTM (Proposed)** | **52.73** | **75.73** |
| | No cross-domain [103] | 94.14 | 172.70 |
| | Incorporating social dataset [103] | 91.89 | 166.18 |
| | Incorporating BSs dataset [103] | 93.85 | 167.75 |
| | Incorporating POIs dataset [103] | 90.95 | 164.31 |
| | No transferring [103] | 111.78 | 186.12 |
| | Transferring with sms [103] | 97.82 | 168.87 |
| | Transferring with call [103] | 94.34 | 169.53 |

Table 4.16: Comparison between models in predicting call traffic

| Traffic | Model | MAE | RMSE |
|---|---|---|---|
| **Call** | Spatiotemporal baseline | 15.23 | 22.06 |
| | **2D-ConvLSTM (Proposed)** | **8.98** | **15.02** |
| | No cross-domain [103] | 17.74 | 40.11 |
| | Incorporating social dataset [103] | 18.00 | 37.04 |
| | Incorporating BSs dataset [103] | 17.70 | 33.83 |
| | Incorporating POIs dataset [103] | 15.85 | 33.34 |
| | No transferring [103] | 16.87 | 35.43 |
| | Transferring with sms [103] | 15.72 | 33.47 |
| | Transferring with internet [103] | 14.42 | 30.85 |

Table 4.17: Comparison between models in predicting sms traffic

| Traffic | Model | MAE | RMSE |
|---|---|---|---|
| **Sms** | Spatiotemporal baseline | 24.03 | 36.04 |
| | **2D-ConvLSTM (Proposed)** | **14.42** | **26.60** |
| | No cross-domain [103] | 32.60 | 57.71 |
| | Incorporating social dataset [103] | 27.31 | 55.59 |
| | Incorporating BSs dataset [103] | 28.74 | 54.52 |
| | Incorporating POIs dataset [103] | 28.17 | 52.88 |
| | No transferring [103] | 28.32 | 55.07 |
| | Transferring with call [103] | 25.90 | 50.96 |
| | Transferring with internet [103] | 25.41 | 52.77 |

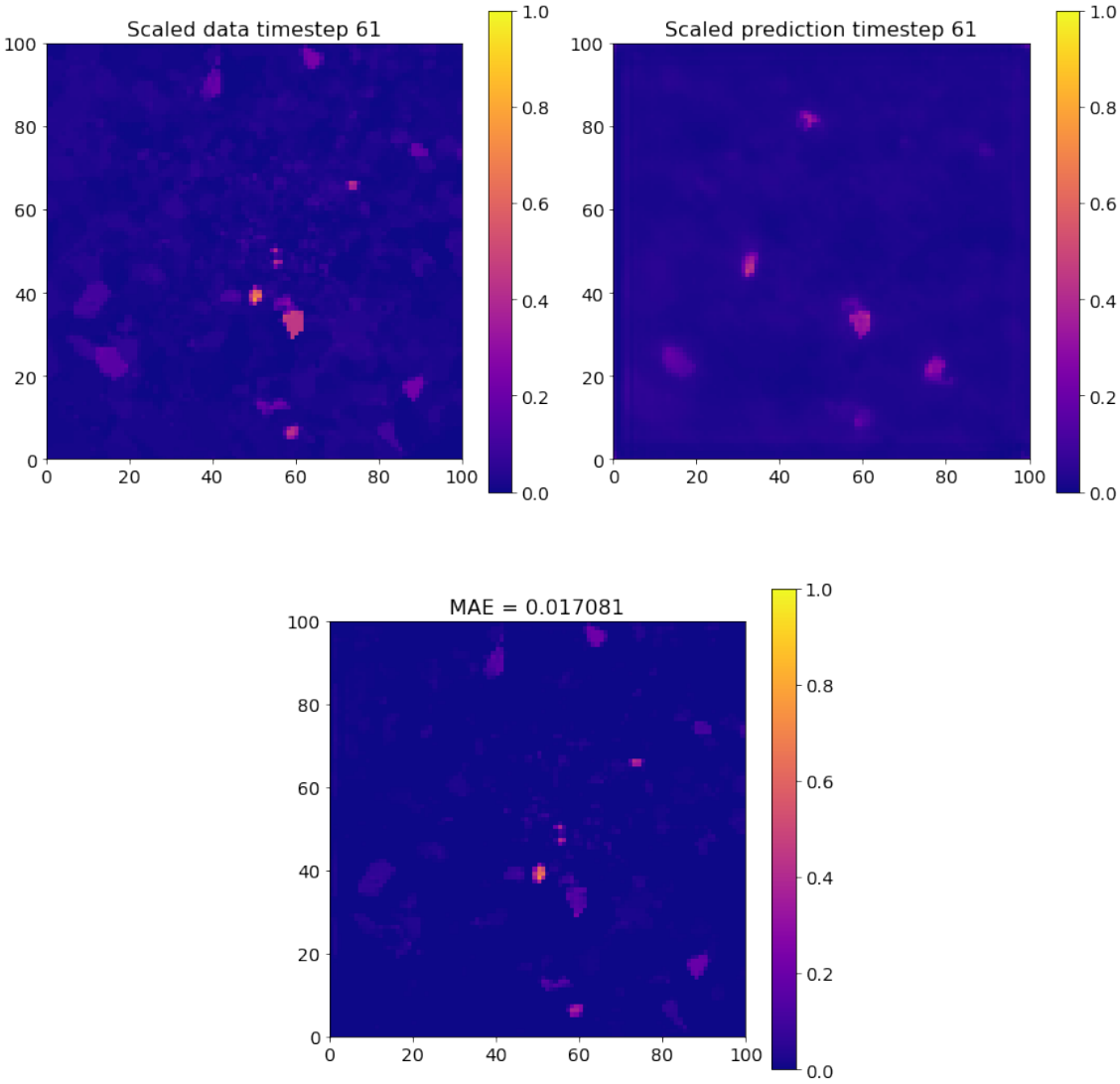


Figure 4.30: 2D-ConvLSTM sms prediction performance on time step 61 .

best prediction results in terms of RMSE and MAE for all three kinds of cellular traffic including internet, call, and, sms as illustrated in Table 4.15 - Table 4.17. The reason is the multivariate nature of our model which incorporates the correlation among variables into the prediction. Also, rather than sticking with the same filter size of 16 in different layers as in STCNet, in our framework, the number of channels were reduced progressively in the four 2D-ConvLSTM layers which optimizes the model size and memory allocation. Therefore, the complexity and execution time of our proposed model as depicted in Table 4.14 are relatively low for predicting the internet, call, and sms traffic.

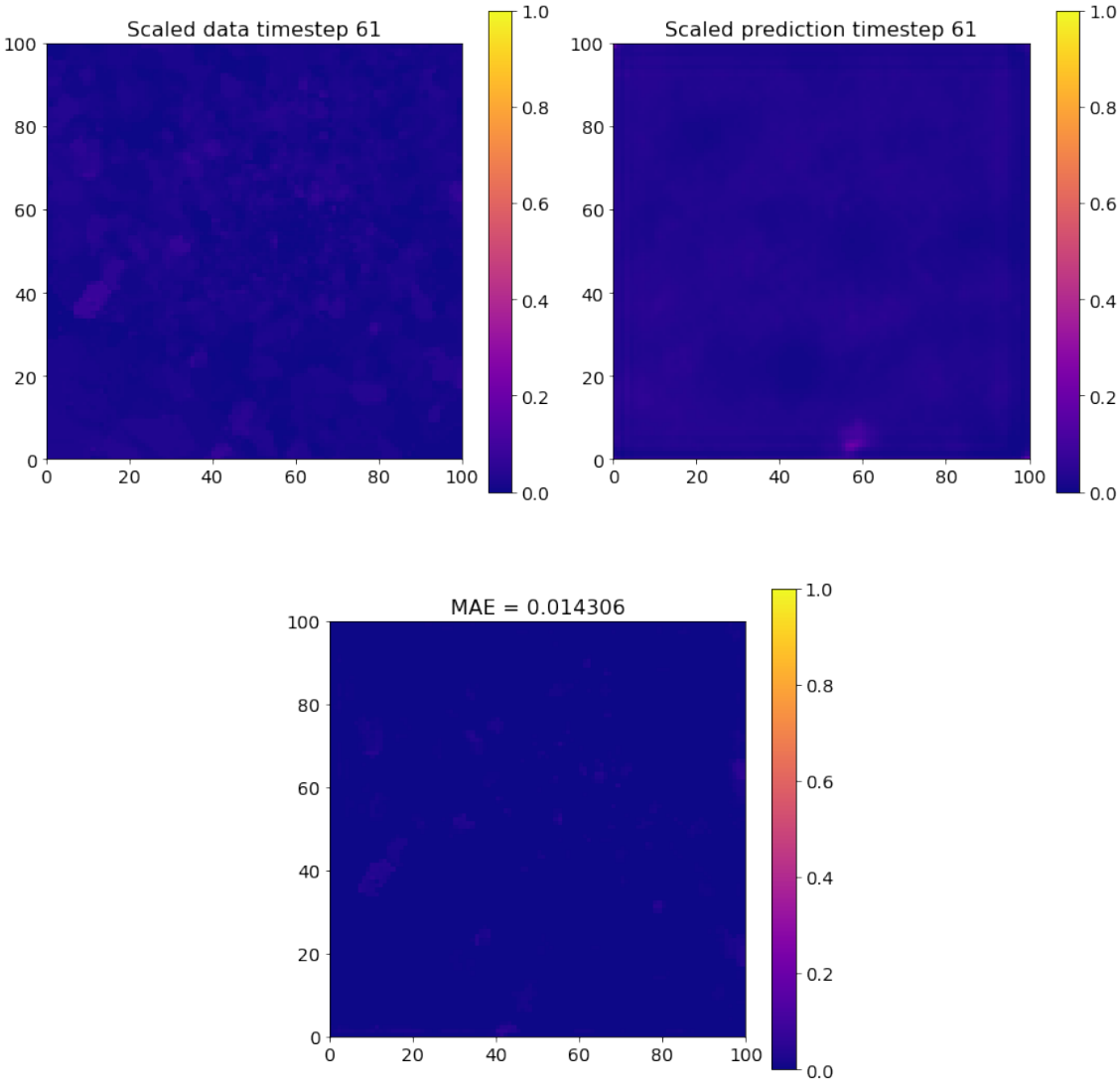


Figure 4.31: 2D-ConvLSTM call prediction performance on time step 61 .

Another factor which contribute to the success of our model can be deploying a new measure named “count” along with “sms”, “call”, and “internet” as the model input. Utilizing count helps to predict various types of cellular traffic more accurately since this measure indicates the number of records in a specific time for a particular grid id.

## 4.6 Conclusion

To anticipate cellular traffic, temporal and spatiotemporal frameworks are developed in this study. Among temporal models, FCSN and 1D-CNN have comparable performance with the

smallest MAE. However, 1D-CNN is a smaller network with less number of parameters. Moreover, the proposed 1D-CNN is having less complexity and shows smaller execution time for predicting the next 24-hour traffic. In terms of spatiotemporal prediction, the proposed 2D-ConvLSTM model showed better performance for predicting all three kinds of cellular traffic, including internet, sms, and calls compared to temporal analysis which showed the effectiveness of incorporating data dependencies in traffic prediction. It is crucial to recognize the resource requirements for each slice and how these requirements vary over time. Accurate predictions of the models would help in avoiding under-provisioning which leads to poor network slice performance and poor QoS for users. Also, over-provisioning can incur expenditures on infrastructure providers. Hence, as dynamically adjusting the resource allocation to network slices in 5G network is required, forecasting the traffic profiles of each slice is vital.

# Chapter 5

# Conclusion

## 5.1 Discussion and Conclusion

Due to the continuously growing various kinds of cellular traffic, mobile traffic forecasting is becoming a significant aspect of the optimization of cellular networks. Accurately anticipating internet traffic can help with resource allocation in 5G networks by devoting the precise necessary resources for each slice which decreases service provider costs and enhances the performance of network slices. In this study, we focused on predicting the cellular traffic of the next 24 hours by utilizing temporal and spatiotemporal approaches and the prediction performance of several neural network models was assessed. The temporal framework specifically consists of the FCSN, 1D-CNN, SS-LSTM, and AR-LSTM. For the spatiotemporal framework, a 2D-ConvLSTM model was proposed to predict cellular traffic for the upcoming 24 hours. For both frameworks, baseline models were also provided to evaluate the performance of the neural network models.

The results indicate that the FCSN and 1D-CNN have comparable performance in temporal models. However, SS-LSTM model outperforms other temporal models in terms of both MAE and RMSE metrics when it comes to the call traffic forecasting. Regarding sms traffic, SS-LSTM has the lowest MAE while the FCSN better preformed in terms of RMSE. Moreover, 1D-CNN surpasses other temporal models for forecasting internet traffic in terms of MAE and

RMSE. Also, it should be mentioned that 1D-CNN is a smaller network with less number of parameters and it is adequate for edge deployment. Moreover, the proposed 1D-CNN is less complicated and has a faster execution time for predicting the next 24-hour traffic.

Furthermore, the 2-dimensional Convolutional LSTM (2D-ConvLSTM) network was proposed to forecast the individual cellular traffic such as the internet, sms, and calls using a multichannel spatiotemporal data as an input. The prediction accuracy is improved by this multivariate spatiotemporal analysis, due to the capability of extracting the dependencies among variables, spatial and temporal data characteristics and incorporating that into the prediction. The model can anticipate the next 24-hour traffic of internet, sms, and call with an RMSE value of 75.73, 26.60, and 15.02 and an MAE of 52.73, 14.42, and 8.98 respectively, after 24-hour observations of historical data. The 2D-ConvLSTM model was also optimized for memory efficiency by shrinking the width progressively and reducing the number of channels in each layers.

Overall, the experimental results demonstrate that the proposed spatiotemporal model outperforms the temporal models and other techniques in the literature in forecasting the cellular traffic including internet, sms, and call. Hence, it is expected that more effective network optimization and resource allocation would be possible by predicting the cellular traffic via proposed 2D-ConvLSTM model. In future work, the prediction performance can be improved and deployment of the 2D-ConvLSTM model in 5G networks can be optimized automatically. The cost incurred by MNOs for allocating resources to each slice before and after using the 2D-ConvLSTM model for cellular traffic forecasting can be also calculated and compared. Future research will also leverage other datasets to measure the model generalization capability of the proposed spatiotemporal 2D-ConvLSTM model. Another potential next step is to use more historical data that can be used to forecast different types of cellular traffic. For instance, by having the whole year data of internet, sms, and calls, yearly and monthly traffic patterns can be captured which can be used by the 2D-ConvLSTM model to predict the cellular traffic for the next months and years and even predict the traffic of holidays during each year. The cellular traffic prediction of holidays would be helpful for MNOs in terms of appropriate re-

source allocation before different holidays each year. Also, the prediction capability of the proposed models over longer time frames will be tested in future works. Moreover, Model pruning and optimization for edge deployment can be explored as well. Another possible future work would be conducting statistical analysis to identify extreme values and considering how the proposed models can handle outliers. Finally, one beneficial further step is considering the effect of network performance such as latency in the cellular traffic prediction.

# Bibliography


[1] U Cisco. Cisco annual internet report (2018–2023) white paper. *Online](accessed March 26, 2021) https://www. cisco. com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/whitepaper-c11-741490. html*, 2020.

[2] Alireza Jolfaei, Krishna Kant, and Hassan Shafei. Secure data streaming to untrusted road side units in intelligent transportation system. In *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pages 793–798. IEEE, 2019.

[3] Gianni Barlacchi, Marco De Nadai, Roberto Larcher, Antonio Casella, Cristiana Chitic, Giovanni Torrisi, Fabrizio Antonelli, Alessandro Vespignani, Alex Pentland, and Bruno Lepri. A multi-source dataset of urban life in the city of milan and the province of trentino. *Scientific data*, 2(1):1–15, 2015.

[4] Dynamically reconfigurable optical-wireless backhaul/fronthaul with cognitive control plane for small cells and cloud-rans. https://www.5g-xhaul-project.eu/project_motivation.html. Accessed: 2022-1-18.

[5] Muhammad Ali Imran, Yusuf Abdulrahman Sambo, and Qammer H Abbasi. *Enabling 5G communication systems to support vertical industries*. John Wiley & Sons, 2019.

[6] NGMN Alliance. Ngmn 5g white paper, 2015.

[7] Xin Li, Mohammed Samaka, H Anthony Chan, Deval Bhamare, Lav Gupta, Chengcheng Guo, and Raj Jain. Network slicing for 5g: Challenges and opportunities. *IEEE Internet Computing*, 21(5):20–27, 2017.

[8] 3GPP TS 23.503, 'Policy and charging control framework for the 5G System (5GS), \OT1\textquotedblrighthttps://www.etsi.org/deliver/etsi_ts/123500123599/123503/15.08.00_60/ts_123503v150800p.pdf.

[9] Prakash Suthar, Vivek Agarwal, Rajaneesh Sudhakar Shetty, and Anil Jangam. Migration and interworking between 4g and 5g. In *2020 IEEE 3rd 5G World Forum (5GWF)*, pages 401–406, 2020.

[10] Rashid Mijumbi, Joan Serrat, Juan-Luis Gorricho, Niels Bouten, Filip De Turck, and Raouf Boutaba. Network function virtualization: State-of-the-art and research challenges. *IEEE Communications Surveys Tutorials*, 18(1):236–262, 2016.

[11] ETSI GS NFV-MAN 001 V1.1.1: Network Functions Virtualisation (NFV); Management and orchestration, ETSI Ind. Spec. Group (ISG) Netw. Functions Virtualisation (NFV),' 2014, page 1-183.

[12] Maede Zolanvari. Sdn for 5g. *Cse. Wustl. Edu*, 2015.

[13] Giuseppe A Carella, Michael Pauls, Thomas Magedanz, Marco Cilloni, Paolo Bellavista, and Luca Foschini. Prototyping nfv-based multi-access edge computing in 5g ready networks with open baton. In *2017 IEEE Conference on Network Softwarization (NetSoft)*, pages 1–4. IEEE, 2017.

[14] Tuyen X. Tran, Abolfazl Hajisami, Parul Pandey, and Dario Pompili. Collaborative mobile edge computing in 5g networks: New paradigms, scenarios, and challenges. *IEEE Communications Magazine*, 55(4):54–61, 2017.

[15] Vincenzo Sciancalepore, Fabio Giust, Konstantinos Samdanis, and Zarrar Yousaf. A double-tier mec-nfv architecture: Design and optimisation. In *2016 IEEE Conference on standards for communications and networking (CSCN)*, pages 1–6. IEEE, 2016.

[16] Yun Chao Hu, Milan Patel, Dario Sabella, Nurit Sprecher, and Valerie Young. Mobile edge computing—a key technology towards 5g. *ETSI white paper*, 11(11):1–16, 2015.

[17] Mohd Abdul Ahad, Sara Paiva, Gautami Tripathi, and Noushaba Feroz. Enabling technologies and sustainable smart cities. *Sustainable cities and society*, 61:102301, 2020.

[18] Timeline for deployment of c-v2x. https://5gaa.org/wp-content/uploads/2019/01/5GAA_White-Paper-CV2X-Roadmap.pdf. Accessed: 2022-1-12.

[19] V2i system. https://www.its.dot.gov/v2i/. Accessed: 2022-1-12.

[20] vehicle-to-vehicle-communication. https://www.nhtsa.gov/technology-innovation/vehicle-vehicle-communication. Accessed: 2022-1-12.

[21] vehicle-to-everything(v2x). https://corporatefinanceinstitute.com/resources/knowledge/other/vehicle-to-everything-v2x/. Accessed: 2022-1-12.

[22] Ali Gohar and Gianfranco Nencioni. The role of 5g technologies in a smart city: The case for intelligent transportation system. *Sustainability*, 13(9):5188, 2021.

[23] Maria Rita Palattella, Mischa Dohler, Alfredo Grieco, Gianluca Rizzo, Johan Torsner, Thomas Engel, and Latif Ladid. Internet of things in the 5g era: Enablers, architecture, and business models. *IEEE journal on selected areas in communications*, 34(3):510–527, 2016.

[24] Debasis Bandyopadhyay and Jaydip Sen. Internet of things: Applications and challenges in technology and standardization. *Wireless personal communications*, 58(1):49–69, 2011.

[25] Arun Narayanan, Arthur Sousa De Sena, Daniel Gutierrez-Rojas, Dick Carrillo Melgarejo, Hafiz Majid Hussain, Mehar Ullah, Suzan Bayhan, and Pedro HJ Nardelli. Key advances in pervasive edge computing for industrial internet of things in 5g and beyond. *IEEE Access*, 8:206734–206754, 2020.

[26] A Alter, P Banerjee, PE Daugherty, and W Negm. Driving unconventional growth through the industrial internet of things, 2014.

[27] Petar Popovski, Kasper Fløe Trillingsgaard, Osvaldo Simeone, and Giuseppe Durisi. 5g wireless network slicing for embb, urllc, and mmtc: A communication-theoretic view. *Ieee Access*, 6:55765–55779, 2018.

[28] Shahid Mumtaz, Ai Bo, Anwer Al-Dulaimi, and Kim-Fung Tsang. Guest editorial 5g and beyond mobile technologies and applications for industrial iot (iiot). *IEEE Transactions on Industrial Informatics*, 14(6):2588–2591, 2018.

[29] Yushan Siriwardhana, Pawani Porambage, Madhusanka Liyanage, Jaspreet Singh Walia, Marja Matinmikko-Blue, and Mika Ylianttila. Micro-operator driven local 5g network architecture for industrial internet. In *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–8. IEEE, 2019.

[30] Conor Sexton, Quentin Bodinier, Arman Farhang, Nicola Marchetti, Faouzi Bader, and Luiz A DaSilva. Enabling asynchronous machine-type d2d communication using multiple waveforms in 5g. *IEEE internet of things journal*, 5(2):1307–1322, 2018.

[31] A slice in time: Slicing security in 5g core networks. https://info.adaptivemobile.com/5g-network-slicing-security. Accessed: 2021-10-26.

[32] Dimitrios Michael Manias and Abdallah Shami. The need for advanced intelligence in nfv management and orchestration. *IEEE Network*, 35(1):365–371, 2020.

[33] Will 5g complexity overwhelm mnos? https://www.guavus.com/will-5g-complexity-overwhelm-mnos/. Accessed: 2021-10-26.

[34] Is 5g technology dangerous? - pros and cons of 5g network. https://www.kaspersky.com/resource-center/threats/5g-pros-and-cons. Accessed: 2021-10-26.

[35] Salman Rashid and Shukor Abd Razak. Big data challenges in 5g networks. In *2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN)*, pages 152–157. IEEE, 2019.

[36] Parthasarathy Guturu. Explosive wireless consumer demand for network bandwidth-fifth generation and beyond [future directions]. *IEEE consumer electronics magazine*, 6(2):27–31, 2017.

[37] Osianoh Glenn Aliu, Ali Imran, Muhammad Ali Imran, and Barry Evans. A survey of self organisation in future cellular networks. *IEEE Communications Surveys & Tutorials*, 15(1):336–361, 2012.

[38] Tamer Omar, Thomas Ketseoglou, and Ibrahim Naffaa. A novel self-healing model using precoding & big-data based approach for 5g networks. *Pervasive and Mobile Computing*, 73:101365, 2021.

[39] 3GPP TS 132541, 'telecommunications system (phase 2+); universal mobile telecommunications system (UMTS); lte; telecommunication management; self-organizing networks (SON); self-healing concepts and requirements, version 11 release 11,' 2012, https://www.etsi.org/deliver/etsi_ts.

[40] Péter Szilágyi and Szabolcs Nováczki. An automatic detection and diagnosis framework for mobile communication systems. *IEEE transactions on Network and Service Management*, 9(2):184–197, 2012.

[41] Yuhong Huang, Xiaodong Xu, Nan Li, Haiyu Ding, and Xiaoxuan Tang. Prospect of 5g intelligent networks. *IEEE Wireless Communications*, 27(4):4–5, 2020.

[42] Wei Jiang, Mathias Strufe, and Hans D Schotten. Experimental results for artificial intelligence-based self-organized 5g networks. In *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pages 1–6. IEEE, 2017.

[43] Pedro Neves, Rui Calé, Mário Rui Costa, Carlos Parada, Bruno Parreira, Jose Alcaraz-Calero, Qi Wang, James Nightingale, Enrique Chirivella-Perez, Wei Jiang, et al. The selfnet approach for autonomic management in an nfv/sdn networking paradigm. *International Journal of Distributed Sensor Networks*, 12(2):2897479, 2016.

[44] Focus group on technologies for network 2030. https://www.itu.int/en/ITU-T/focusgroups/net2030/Pages/default.aspx. Accessed: 2022-01-25.

[45] 6g flagship, univ. oulu, oulu, finland,. https://www.oulu.fi/6gflagship/. Accessed: 2022-01-25.

[46] Madhusanka Liyanage, Andrei Gurtov, and Mika Ylianttila. *Software defined mobile networks (SDMN): beyond LTE network architecture*. John Wiley & Sons, 2015.

[47] Walid Saad, Mehdi Bennis, and Mingzhe Chen. A vision of 6g wireless systems: Applications, trends, technologies, and open research problems. *IEEE network*, 34(3):134–142, 2019.

[48] Fang Fang, Yanqing Xu, Quoc-Viet Pham, and Zhiguo Ding. Energy-efficient design of irs-noma networks. *IEEE Transactions on Vehicular Technology*, 69(11):14088–14092, 2020.

[49] Yang Lu and Xianrong Zheng. 6g: A survey on technologies, scenarios, challenges, and the related issues. *Journal of Industrial Information Integration*, page 100158, 2020.

[50] Yi Liu, Xingliang Yuan, Zehui Xiong, Jiawen Kang, Xiaofei Wang, and Dusit Niyato. Federated learning for 6g communications: Challenges, methods, and future directions. *China Communications*, 17(9):105–118, 2020.

[51] vepc in lte networks: Time to move ahead. https://www.nokia.com/networks/solutions/cloud-packet-core/. Accessed: 2022-01-27.

[52] Maryam Imran Sheik Mamode and Tulsi Pawan Fowdur. Survey of scheduling schemes in 5g mobile communication systems. *Journal of Electrical Engineering, Electronics, Control and Computer Science*, 6(2):21–30, 2020.

[53] Stefan Schröder. Security in 5g inter-network signalling. https://docbox.etsi.org/Workshop/2018/201806_ETSISECURITYWEEK/5G/S02_SECURITY_5G_INTER-NWK_SIGNALLING_/SBA_INTRO_SA3_STATUS_T-SYSTEMS_SCHROEDER.pdf, 2015. Accessed: 2022–01-27.

[54] Konstantinos Samdanis, Athul Prasad, Min Chen, and Kai Hwang. Enabling 5g verticals and services through network softwarization and slicing. *IEEE Communications Standards Magazine*, 2(1):20–21, 2018.

[55] Silvia Sekander, Hina Tabassum, and Ekram Hossain. Multi-tier drone architecture for 5g/b5g cellular networks: Challenges, trends, and prospects. *IEEE Communications Magazine*, 56(3):96–103, 2018.

[56] NGMN Alliance. Service-based architecture in 5g: case study and deployment recommendations, 2019.

[57] Ivezis, m. 2020 introduction to 5g core service-based architecture (sba). https://5g.security/5g-technology/5g-core-sba-components-architecture/. Accessed: 2022-02-04.

[58] Gabrial Brown. Service-based architecture for 5g core networks. *Huawei White Paper*, 1, 2017.

[59] 5g core network – architecture, network functions, and interworking. https://www.rfglobalnet.com/doc/g-core-network-architecture-network-functions-and-interworking-0001. Accessed: 2022-02-07.

[60] Young-il Choi and Noik Park. Slice architecture for 5g core network. In *2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN)*, pages 571–575, 2017.

[61] Standalone 5g vs. non-standalone 5g. https://www.rcrwireless.com/20210907/5g/standalone-5g-vs-non-standalone-5g. Accessed: 2022-02-10.

[62] What is non-standalone 5g? https://www.everythingrf.com/community/what-is-non-standalone-5g. Accessed: 2022-02-10.

[63] 5g network introduction & migration paths. https://www.gsma.com/futurenetworks/wp-content/uploads/2018/04/Road-to-5G-Introduction-and-Migration_FINAL.pdf. Accessed: 2022-02-11.

[64] 5g migration strategy from eps to 5g system. https://www.ericsson.com/493cea/assets/local/reports-papers/ericsson-technology-review/docs/2020/migration-from-eps-to-5g.pdf. Accessed: 2022-02-11.

[65] 3GPP TS 23.501, ‘ System architecture for the 5G System (5GS), \OT1\textquotedblrighthttps://www.etsi.org/deliver/etsi_ts123500_123599/123501/15.09.00_60/ts_123501v150900p.pdf.

[66] Xiaofei Wang, Xiuhua Li, and Victor CM Leung. Artificial intelligence-based techniques for emerging heterogeneous network: State of the arts, opportunities, and challenges. *IEEE Access*, 3:1379–1391, 2015.

[67] Jong-Hyouk Lee and Hyoungshick Kim. Security and privacy challenges in the internet of things [security and privacy matters]. *IEEE Consumer Electronics Magazine*, 6(3):134–136, 2017.

[68] Sutapa Sarkar and Aritri Debnath. Machine learning for 5g and beyond: Applications and future directions. In *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 1688–1693. IEEE, 2021.

[69] Taiwo Oladipupo Ayodele. Types of machine learning algorithms. In Yagang Zhang, editor, *New Advances in Machine Learning*, chapter 3. IntechOpen, Rijeka, 2010.

[70] Manuel Eugenio Morocho-Cayamcela, Haeyoung Lee, and Wansu Lim. Machine learning for 5g/b5g mobile and wireless communications: Potential, limitations, and future directions. *IEEE Access*, 7:137184–137206, 2019.

[71] Jasneet Kaur, M Arif Khan, Mohsin Iftikhar, Muhammad Imran, and Qazi Emad Ul Haq. Machine learning techniques for 5g and beyond. *IEEE Access*, 9:23472–23488, 2021.

[72] Nasir Abbas, Yan Zhang, Amir Taherkordi, and Tor Skeie. Mobile edge computing: A survey. *IEEE Internet of Things Journal*, 5(1):450–465, 2018.

[73] Keith Shaw. What is edge computing and why it matters. *Network World*, 2019.

[74] Erik Nygren, Ramesh K Sitaraman, and Jennifer Sun. The akamai network: a platform for high-performance internet applications. *ACM SIGOPS Operating Systems Review*, 44(3):2–19, 2010.

[75] Ning Wang, Ekram Hossain, and Vijay K Bhargava. Backhauling 5g small cells: A radio resource management perspective. *IEEE Wireless Communications*, 22(5):41–49, 2015.

[76] Fabio Giust, Luca Cominardi, and Carlos J Bernardos. Distributed mobility management for future 5g networks: overview and analysis of existing approaches. *IEEE Communications Magazine*, 53(1):142–149, 2015.

[77] Emad Aqeeli, Abdallah Moubayed, and Abdallah Shami. Power-aware optimized rrh to bbu allocation in c-ran. *IEEE Transactions on Wireless Communications*, 17(2):1311–1322, 2017.

[78] Mamta Agiwal, Abhishek Roy, and Navrati Saxena. Next generation 5g wireless networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 18(3):1617–1655, 2016.

[79] Akhil Gupta and Rakesh Kumar Jha. A survey of 5g network: Architecture and emerging technologies. *IEEE access*, 3:1206–1232, 2015.

[80] Hassan Hawilo, Manar Jammal, and Abdallah Shami. Network function virtualization-aware orchestrator for service function chaining placement in the cloud. *IEEE Journal on Selected Areas in Communications*, 37(3):643–655, 2019.

[81] Dario Bega, Marco Gramaglia, Marco Fiore, Albert Banchs, and Xavier Costa-Perez. Deepcog: Cognitive network management in sliced 5g networks with deep learning. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 280–288. IEEE, 2019.

[82] Andrea Fendt, Simon Lohmuller, Lars Christoph Schmelz, and Bernhard Bauer. A network slice resource allocation and optimization model for end-to-end mobile networks. In *2018 IEEE 5G World Forum (5GWF)*, pages 262–267. IEEE, 2018.

[83] Ali Imran, Ahmed Zoha, and Adnan Abu-Dayya. Challenges in 5g: how to empower son with big data for enabling 5g. *IEEE network*, 28(6):27–33, 2014.

[84] Fengli Xu, Yuyun Lin, Jiaxin Huang, Di Wu, Hongzhi Shi, Jeungeun Song, and Yong Li. Big data driven mobile traffic understanding and forecasting: A time series approach. *IEEE transactions on services computing*, 9(5):796–805, 2016.

[85] Kan Zheng, Zhe Yang, Kuan Zhang, Periklis Chatzimisios, Kan Yang, and Wei Xiang. Big data-driven optimization for mobile networks toward 5g. *IEEE network*, 30(1):44–51, 2016.

[86] Chunxiao Jiang, Haijun Zhang, Yong Ren, Zhu Han, Kwang-Cheng Chen, and Lajos Hanzo. Machine learning paradigms for next-generation wireless networks. *IEEE Wireless Communications*, 24(2):98–105, 2016.

[87] Duong D Nguyen, Hung X Nguyen, and Langford B White. Reinforcement learning with network-assisted feedback for heterogeneous rat selection. *IEEE Transactions on Wireless Communications*, 16(9):6062–6076, 2017.

[88] Fairuz Amalina Narudin, Ali Feizollah, Nor Badrul Anuar, and Abdullah Gani. Evaluation of machine learning classifiers for mobile malware detection. *Soft Computing*, 20(1):343–357, 2016.

[89] Mohammad Abu Alsheikh, Dusit Niyato, Shaowei Lin, Hwee-Pink Tan, and Zhu Han. Mobile big data analytics using deep learning and apache spark. *IEEE network*, 30(3):22–29, 2016.

[90] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[91] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[92] Tiago Prado Oliveira, Jamil Salem Barbar, and Alexsandro Santos Soares. Computer network traffic prediction: a comparison between traditional and deep learning neural networks. *International Journal of Big Data Intelligence*, 3(1):28–37, 2016.

[93] Jing Wang, Jian Tang, Zhiyuan Xu, Yanzhi Wang, Guoliang Xue, Xing Zhang, and Dejun Yang. Spatiotemporal modeling and prediction in cellular networks: A big data enabled deep learning approach. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, pages 1–9. IEEE, 2017.

[94] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[95] Chuanting Zhang, Haixia Zhang, Dongfeng Yuan, and Minggao Zhang. Citywide cellular traffic prediction based on densely connected convolutional neural networks. *IEEE Communications Letters*, 22(8):1656–1659, 2018.

[96] Jiansheng Lin, Youjia Chen, Haifeng Zheng, Ming Ding, Peng Cheng, and Lajos Hanzo. A data-driven base station sleeping strategy based on traffic prediction. *IEEE Transactions on Network Science and Engineering*, 2021.

[97] Wei-Che Chien and Yueh-Min Huang. A lightweight model with spatial–temporal correlation for cellular traffic prediction in internet of things. *The Journal of Supercomputing*, 77(9):10023–10039, 2021.

[98] Qingyao Liu, Jianwu Li, and Zhaoming Lu. St-tran: Spatial-temporal transformer for cellular traffic prediction. *IEEE Communications Letters*, 25(10):3325–3329, 2021.

[99] Yin Gao, Man Zhang, Jiajun Chen, Jiren Han, Dapeng Li, and Ruitao Qiu. Accurate load prediction algorithms assisted with machine learning for network traffic. In *2021 International Wireless Communications and Mobile Computing (IWCMC)*, pages 1683–1688. IEEE, 2021.

[100] Xu Zhou, Yong Zhang, Zhao Li, Xing Wang, Juan Zhao, and Zhao Zhang. Large-scale cellular traffic prediction based on graph convolutional networks with transfer learning. *Neural Computing and Applications*, pages 1–11, 2022.

[101] Nan Zhao, Aonan Wu, Yiyang Pei, Ying-Chang Liang, and Dusit Niyato. Spatial-temporal aggregation graph convolution network for efficient mobile cellular traffic prediction. *IEEE Communications Letters*, 2021.

[102] Chengsheng Pan, Jiang Zhu, Zhixiang Kong, Huaifeng Shi, and Wensheng Yang. Dc-stgcn: dual-channel based graph convolutional networks for network traffic forecasting. *Electronics*, 10(9):1014, 2021.

[103] Chuanting Zhang, Haixia Zhang, Jingping Qiao, Dongfeng Yuan, and Minggao Zhang. Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data. *IEEE Journal on Selected Areas in Communications*, 37(6):1389–1401, 2019.

[104] Nan Zhao, Zhiyang Ye, Yiyang Pei, Ying-Chang Liang, and Dusit Niyato. Spatial-temporal attention-convolution network for citywide cellular traffic prediction. *IEEE Communications Letters*, 24(11):2532–2536, 2020.

[105] Qingtian Zeng, Qiang Sun, Geng Chen, Hua Duan, Chao Li, and Ge Song. Traffic prediction of wireless cellular networks based on deep transfer learning and cross-domain data. *IEEE Access*, 8:172387–172397, 2020.

[106] Chapter 4. fully connected deep networks. https://www.oreilly.com/library/view/tensorflow-for-deep/9781491980446/ch04.html. Accessed: 2022-06-13.

[107] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

# Curriculum Vitae

| | |
|---|---|
| **Name:** | Maryam Mohseni |
| **Post-Secondary Education and Degrees:** | Islamic Azad University South Tehran Branch,<br>Tehran, Iran<br>2015 - 2019<br>B.E. in Computer Engineering – Computer Systems Architecture<br><br>The University of Western Ontario<br>London, Ontario, Canada<br>2021 - 2022<br>M.E.Sc. in Software Engineering |
| **Related Work Experience:** | Teaching Assistant<br>The University of Western Ontario<br>2021 - 2022<br><br>Research Assistant<br>The University of Western Ontario<br>2021 - 2022 |

**Publications:**

M. Mohseni, S. Nikan and A. Shami, ”AI-Based Traffic Forecasting in 5G Network,” 2022 35th Canadian Conference on Electrical and Computer Engineering(CCECE), 2022.

Bolhasani, H., Mohseni, M. & Rahmani, A. M. (2021), ‘Deep learning applications for iot in health care: Asystematic review’,Informatics in Medicine Unlocked23, 100550.URL:https://www.sciencedirect.com/science/article/pii/S235291482100040X