Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

Telecommunication Network Engineering Graduate Program

A Deep Neural Network Approach for Spatiotemporal Throughput Prediction in LTE Network

By

Redamichael Jemberu

Advisor

Dr. -Ing. Dereje Hailemariam

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Telecommunication Network Engineering

October 2021

Addis Ababa, Ethiopia

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

Telecommunication Network Engineering Graduate Program

A Deep Neural Network Approach for Spatiotemporal Throughput Prediction in LTE Network

By: Redamichael Jemberu

Approval by Board of Examiners

_____	_____	__ / /2021
Chairman, School Graduate Committee	Signature	Date

Committee

Dr. -Ing. Dereje Hailemariam	_____	/ /2021
Advisor	Signature	Date
_____	_____	/ /2021
Internal examiner	Signature	Date
_____	_____	/ /2021
External examiner	Signature	Date

Declaration

I, the undersigned, declare that this thesis is my original work and does not incorporate without recognition of any material previously submitted for a degree or diploma in any other university or institute of higher learning. To the best of my knowledge, it does not contain any material previously published or written by another person unless mentioned in the text.

Redamichael Jemberu _____

Name Signature

Addis Ababa _____/____/2021

Place Date of Submission

This thesis has been submitted for examination with my approval as a university advisor.

Dr. -Ing. Dereje Hailemariam _____

Name Signature

Abstract

In the mobile broadband era, whereby high-resolution video services are emerging, user throughput is a key metric to ensure the satisfaction of user experience. Monitoring and assuring the desired level of user throughput is thus a major concern of mobile network operators (MNOs). Since different services and areas have different bandwidth requirements, it is difficult to guarantee a good user experience using traditional Key Performance Indicator (KPI) based network monitoring and analysis.

To that end, a proactive approach that takes into account spatiotemporal dimensions is required. The enhancement of Deep Neural Network (DNN) algorithms and MNO's big data can be leveraged to solve spatiotemporal modeling problems. Due to its ability to capture spatial patterns, Convolutional Long Short-Term Memory (ConvLSTM) is widely used in video prediction and later applied to various spatiotemporal prediction problems. Although MNO's network management system (NMS) can provide cell-level average user throughput data, spatial mapping of NMS data is challenging mainly due to the varying coverage of cells. In this thesis, a drive test-based coverage data analysis is conducted to overcome this challenge. By applying this analysis, cell coverage in a spatial grid of dimension 100m × 100m is derived. Next, a technique that maps Long Term Evolution (LTE) average downlink (DL) user throughput data (obtained from NMS) to each grid is adopted. Finally, the ConvLSTM algorithm is applied to build a model that could predict grid-level user throughput. The developed model provides a 3-hour future DL user throughput prediction with Root Mean Square Error (RMSE) of 2.02, Mean Absolute Error (MAE) of 1.52, and Mean Absolute percentage Error (MAPE) of 14.68.

Keywords: *LTE, DL user throughput, DNN, ConvLSTM, prediction, spatiotemporal*

Acknowledgment

First and foremost, I would like to thank Ethio telecom for allowing me to pursue this valuable graduate program. I would also like to thank Addis Ababa Institute of Technology, School of Electrical and Computer Engineering, for the outstanding courses and lectures; I'm profoundly grateful. Most of all, I can't thank enough my mentor, Dr.-Ing. Dereje Hailemariam. His follow-up and guidance are to the tiny details, which sometimes astound me. His extra effort enabled me to deliver the best out of myself.

All TNE batch three students, thank you for the good vibes and the support you showed. Here, I would like to thank two of my mates, Tamrat Demena and Endale Mare, in particular, for the good times we had; the journey would've been arduous otherwise. I must also extend my sincere gratitude to a good friend and colleague, Yibeltal Addis, for his invaluable assistance during data gathering.

Thank you, Mom (Ataye), for your unspeakable love and prayers and so my dear sister Woynitu Jemberu for always being there. Finally, I would like to extend my appreciation and gratitude to my friends from outside of school: Minichil Fantahun, Bekalu Alebachew, and Henok Habtu. Your support and understanding were exceptional; Thank you.

Contents

Declaration i

Abstract ii

Acknowledgment iii

List of figures vii

List of Tables ix

Abbreviations, Definitions, Acronym x

1 Introduction 1

1.1 Background 1

1.2 Statement of the Problem 2

1.3 Objective 3

1.3.1 General Objective 3

1.3.2 Specific Objectives 4

1.4 Literature Review 4

1.5 Methodology 7

1.6 Scope and Limitation 7

1.6.1 Scope 7

1.6.2 Limitation 8

1.7 Contribution 8

1.8 Thesis Outline 8

2 Background on LTE Network 9

2.1 LTE Network Overview 9

2.1.1 Evolution towards LTE and Enabling Technologies 9
2.1.2 LTE Network Architecture 11
2.1.3 LTE Protocol Architecture 13
2.1.4 LTE Frame Structure 15
2.2 LTE Network Performance Measurement 16
2.2.1 Throughput Performance Measurement 16
3 Throughput and Traffic Prediciton Techniques 20
3.1 Mobile Data Throughput and Traffic Characteristics 20
3.2 Time Series Forecasting 20
3.2.1 Linear Time Series Models 21
3.2.2 Nonlinear Time Series Models 21
3.2.3 Hybrid Time Series Models 22
3.2.4 Decomposed Time Series Models 22
3.3 Spatial Prediction 22
3.4 Spatiotemporal Prediction 23
3.5 Deep Learning and Mobile Networks 23
3.5.1 Overview of Deep Learning 23
3.5.2 Deep Learning in Mobile Networks 26
4 LTE Throughput Modelling and Prediction 32
4.1 Proposed Approach 32
4.2 Data Collection 33
4.3 Coverage Analysis 33

4.3.1 Handling Null Grids 36
4.4 Spatial Mapping of NMS Data 37
4.5 Data Preprocessing 39
4.5.1 Data Cleaning 39
4.5.2 Data Integration 39
4.5.3 Data Normalization 40
4.6 Model Development 40
4.6.1 Network Architecture 41
4.6.2 Model Training 42
4.6.3 Model Performance Evaluation 44
4.7 Prediction Results 46
4.8 Detecting Poor Coverage 49
4.9 Throughput Distribution Map 50
5 Conclusion and Future Directions 52
5.1 Conclusion 52
5.2 Future Directions 53

List of figures

Figure 2.1 LTE evolution and enabling technologies [12] 11
Figure 2.2 LTE network architecture [13] 12
Figure 2.3 LTE control plane protocol stack [14] 14
Figure 2.4 LTE user plane protocol stack [14] 14
Figure 2.5 LTE FDD frame structure [11] 16
Figure 3.1 Relationship between AI, machine learning, and deep learning [23] 24
Figure 3.2 The architecture of a neural network [23] 25
Figure 3.3 Machine learning life cycle [23] 25
Figure 3.4 Structure of basic CNN [26] 27
Figure 3.5 Operational principle of a 2D convolutional layer [5] 28
Figure 3.6 Structure of a simple RNN [5] 29
Figure 3.7 The inner structure of LSTM cell [28] 30
Figure 4.1 LTE DL user throughput modeling and forecasting system model 33
Figure 4.2 Target area and drive test route 35
Figure 4.3 Drive test measurement samples 35
Figure 4.4 Samples and contributing cells (6 by 6 grid) 35
Figure 4.5 Samples and contributing cells (10 by 10 grid) 35
Figure 4.6 Data integration technique. 40
Figure 4.7 Proposed network architecture 42
Figure 4.8 Training curve for a 12-hour input sequence 44
Figure 4.9 Training curve for a 24-hour input sequence 44
Figure 4.10 Temporal throughput prediction for one cell (24-hour observation) 47
Figure 4.11 Temporal throughput prediction for one cell (12-hour observation) 47
Figure 4.12 MAPE distribution over 100 grids. 48
Figure 4.13 Spatial throughput prediction after 3 hours (24-hour observation) 48

Figure 4.14 Spatial throughput prediction after 3 hours (12-hour observation) 49
Figure 4.15 LTE coverage map with dominant cells..... 50
Figure 4.16 LTE average DL user throughput distribution map 51

List of Tables

Table 2.1 Key features of LTE [10] 10
Table 4.1 Handling of null grids 36
Table 4.2 Hyper parameter tuning space 43
Table 4.3 Best model hyper parameters 43
Table 4.4 Interpretation of MAPE value [9] 46
Table 4.5 Model evaluation results 46

Abbreviations, Definitions, Acronym

3GPP	3rd Generation Partnership Project
4G	Fourth Generation
AI	Artificial Intelligence
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
CDR	Call Detail Record
CNN	Convolutional Neural Network
ConvLSTM2D	Two-dimensional Long-Short Memory
Conv3D	Three-dimensional Convolution
DL	Downlink
DNN	Deep Neural Network
DT	Drive Test
EPC	Evolved Packet Core
EPS	Evolved Packet System
FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
FNN	Feed-Forward Neural Network
HARQ	Hybrid Automatic Repeat Request
HSDPA	High-Speed Downlink Packet Access
HSPA	High-Speed Packet Access
IP	Internet Protocol
KPI	Key Performance Indicator
LSTM	Long-Short Term Memory
LTE	Long Term Evolution
MAC	Medium Access Control

MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MDT	Minimization of Drive Test
MIMO	Multiple-Input Multiple-Output
MME	Mobility Management Entity
MNO	Mobile Network Operator
MSE	Mean Squared Error
NAS	Non-Access Stratum
NMS	Network Management System
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
PDCP	Packet Data Convergence Protocol
QAM	Quadrature Amplitude Modulation
RE	Resource Element
RF	Random Forest
RLC	Radio Link Control
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
RSRP	Reference Signal Received power
SCFDMA	Single Carrier Frequency Division Multiple Access
SDU	Service Data Unit
SVR	Support Vector Regression
TP	Throughput
TTI	Transmission Time Interval
UE	User Equipment
UL	Uplink
UMTS	Universal Mobile Telecommunications System

UTRAN	Universal Terrestrial Radio Access Network
WCDMA	Wideband Code Division Multiple Access

1 Introduction

This chapter discusses the main components of the thesis. It starts with a brief background, followed by conceptualizing the research problem, a concise literature review, and articulating the proposed approach towards solving the identified problem.

1.1 Background

The ongoing demand for high bandwidth and low latency mobile services has led to the continued evolution of network infrastructures; where network densification is one component of the evolution. The consumer side demand and the supply side densification pose a challenge to network management and quality assurance activities [1].

To evaluate the cellular network's quality of service, the Third Generation Partnership Project (3GPP) has defined five classes of service performance indicators; namely, service availability, accessibility, retainability, mobility, and integrity. Throughput is a KPI that falls into the integrity class and gives an insight into the users' experience on the cellular service provider. It's a key factor that may influence the user to change a particular wireless subscription [3]. This makes proactive monitoring and maintaining the agreed level of throughput compelling for service providers.

As part of guaranteeing robust Quality of Experience (QoE), "xMpbs Anytime Anywhere", is becoming a common strategy by vendors for mobile broadband network planning [4]. "xMbps" reflects both cost and user experience, "Anytime" reflects capacity (i.e., network capacity during busy hours in time domain), and "Anywhere" reflects coverage, i.e., spatial domain. This strategy is needed because different services (e.g., web and video), and different areas (e.g., city center and suburban) have different service requirements. This planning approach, however, requires "xMbps" network discovery

(i.e., experience-based mobile broadband visualization). For example, Huawei's mobile broadband visualization solution displays areas with poor user experiences in a 50 m x 50 m grid, thereby helping locate network bottlenecks [4].

Hence, analyzing network performance indicators such as throughput and traffic both in time and space is imperative to guarantee a good QoE and plan a cost-effective network. Recent developments in big data and artificial intelligence (AI) can be leveraged to capture the dynamic nature of throughput and its complex relationships with various factors [3] [5]. Particularly, DNN algorithms such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), which are specialized in modeling spatial and temporal data, can be used to model spatiotemporal network data [5] [6].

To that end, DNN techniques are applied in this thesis to model and predict spatiotemporal DL average user throughput in LTE network.

1.2 Statement of the Problem

As mentioned in Section 1.1, experienced throughput is one of the key QoE metrics in mobile data service. Monitoring and assuring the desired level of user throughput is thus a major concern of MNOs.

For instance, if we look at Ethio Telecom, the incumbent operator in Ethiopia, has employed various tools to monitor and optimize its network. The first is Huawei's Nastar network optimization tool which is employed for performance analysis of its Universal Mobile Telecommunications System (UMTS) network using Minimization of Drive Test (MDT) measurements. The MDT procedure allows operators to collect radio measurements, such as received signal strength and quality with user equipment (UE) location information and a time stamp [7]. However, MDT log files could be very large depending on the configuration, requiring very high storage. Practically the measurement is configured for certain UEs or specific network nodes [7]. The other tools

are network probes (devices attached at different interfaces of the network) which have been deployed in a few selected areas to monitor the quality of service (QoS). Most monitoring and optimization tasks, however, are performed based on KPI monitoring, routinely scheduled drive tests or customer complaint-based diagnostics.

While all of these approaches have their own limitations, they all have one drawback in common; they are not proactive. Hence, a proactive approach needs to be explored.

Another point to consider is that mobile user throughput is subjected to significant fluctuations [8]. This is mainly caused by the dynamic nature of wireless radio channels. The other factor is user mobility which poses a significant variation in the achievable throughput when a UE moves from one cell to another with different characteristics. The throughput fluctuations lead to inefficient use of network capacity and degradation in QoE [8]. Hence, approaches that are capable of capturing the relationship of these dynamic factors should be followed to model and predict user throughput.

To that end, the envisioned work in this thesis applies a DNN technique to effectively model and predict spatiotemporal LTE DL user throughput. Data collected from LTE network deployed in Addis Ababa city, Ethiopia, is used for this research. Findings of the research enable MNOs to detect areas with poor throughput early on before users notice any service degradation.

1.3 Objective

1.3.1 General Objective

The main objective of the thesis is to model and predict LTE DL user throughput variations both in temporal and spatial domains. i.e., spatiotemporally, for a selected area in Addis Ababa city using DNN techniques.

1.3.2 Specific Objectives

The planned specific objectives to accomplish the main goal of this study are:

- Collect LTE average DL user throughput data from NMS;
- Gather LTE coverage-related metrics via drive test;
- Prepare uniform spatial grids of resolution below cell coverage area;
- Apply ConvLSTM DNN algorithm to model and predict grid-level throughput;
- Generate a throughput distribution map of the target area;
- Geolocate areas with poor coverage;
- Identify the dominant cell in each geo-grid.

1.4 Literature Review

A growing number of researches are being conducted recently by applying deep learning on network-level mobile data for network state prediction and traffic forecasting [5]. Related works on throughput and traffic forecasting are presented in the following paragraphs.

Deep learning was implemented in [8] to build a model that predicts LTE DL user throughput using aggregated data obtained from UE and cell. The UE is set up to download a file, and user throughput data and radio-related metrics, which affect throughput, were logged. Network load data was also collected from cell-level measurements in the NMS simultaneously. The factors affecting throughput include Reference Signal Received Power (RSRP), and Signal-to-Interference-Plus-Noise Ratio (SINR). Afterward, different predictive models were built and compared using baseline random forest (RF), support vector regression (SVR), and deep learning techniques such as feed-forward neural network (FNN), and long short-term memory (LSTM). The author concluded that deep learning methods outperform the baseline lightweight machine

learning techniques. However, the model developed is limited to UEs that stay within a given cell and does not take into account the mobility of the users. A similar approach was followed in [3] to model LTE user and cell throughput from radio-related counters that affect throughput. The authors proved by identifying the most significant features through a dependency test and by applying DNN, a better model can be built. The above two studies show that DNN can be used to effectively model and predict LTE DL user throughput.

A machine learning technique that applies Gaussian process regression and exponential smoothing of time series was applied in [1] to predict LTE network performance at locations or times where measurements are not available using data collected from drive tests. The metrics include geographic coordinates, time, received signal strength (RSS), carrier-to-interference ratio (C/I), and DL user throughput. The authors implemented the Holt-Winters exponential smoothing algorithm for the temporal and Gaussian process regression for spatial prediction of individual metrics. The authors pointed out that other metrics were predicted with better accuracy than throughput. They argued that the reason behind this is that throughput measurement varies irregularly across the test region and a lot of data is needed to train the model better.

Some other interesting works have been conducted on spatiotemporal mobile data traffic predictions. The author in [9] applied ConvLSTM deep learning algorithm to model and predict spatiotemporal mobile data traffic in a selected area of Addis Ababa. First, a geographic grid was prepared considering the locations of sites serving the area. Then site-level hourly data traffic measurement from NMS was spatially mapped to each grid based on the sites falling within them. Finally, a model that predicts grid-level traffic was developed. The prediction results show the model can predict six hours with good accuracy given 24-hours of input. The author argued that ConvLSTM can capture

spatiotemporal correlations effectively, and better results could be obtained if performance data on a grid below site level is made available.

Similar work has been conducted in [6]. The authors argued that there is a similarity between a snapshot of traffic measurements at different regions with a video frame. Hence, they proposed a deep spatiotemporal network (STN) by fusing video prediction algorithms such as ConvLSTM and three-dimensional convolution (Conv3D) with multi-layer perception (MLP) for spatiotemporal traffic prediction. The authors also implemented a double spatio-temporal network (D-STN) architecture with an Ouroboros Training Scheme (OTS). The OTS fine-tunes the neural network, allowing earlier predictions to be fed as inputs when ground truth becomes unavailable while keeping prediction errors small. They argue that this approach allows to minimize multi-step prediction errors and make long-term predictions. Finally, they demonstrated that the implemented approach can predict traffic for 10 hours with reasonable accuracy and better performance than a regular STN, and classical techniques such as autoregressive integrated moving average (ARIMA).

In conclusion, most of the works have demonstrated that deep learning architectures outperform classical methods for throughput and traffic predictions. Furthermore, video prediction algorithms such as ConvLSTM and Conv3D have been proven to offer better performance for spatiotemporal prediction than time-series prediction algorithms such as LSTM.

Many studies on spatiotemporal cellular network prediction have been conducted in the area of data traffic, while throughput has received less attention. This is because throughput is more dynamic by nature, and effectively modeling it requires a lot of measurement data from lower spatial resolution. In spatiotemporal mobile data traffic modeling, site-level traffic data is often mapped spatially using the sites' geolocation since obtaining data below site level is challenging. In this study, an approach that can

generate user throughput data at a grid resolution lower than site-level is used to model and predict LTE DL user throughput.

1.5 Methodology

The first step in conducting this research is to review related works and grasp different techniques that could be used to solve the problem identified as part of this research. Next, a target area is selected based on network activity.

After the target area has been selected, serving sites are identified from network planning data. Next, the necessary data is collected in two ways: from NMS, and from field measurement via drive test. The drive test data is needed to obtain realistic coverage of each cell within the target. To make modeling more convenient, the coverage areas are divided into regular square grids. The data from NMS is then spatially mapped onto each grid based on the coverage probabilities of each cell in the grids.

Finally, a DNN algorithm, specifically ConvLSTM is applied to model and predict spatiotemporal LTE DL user throughput using the Keras deep learning library. Different training schemes and tuning of associated hyper parameters are carried out to obtain a model with optimum results.

1.6 Scope and Limitation

1.6.1 Scope

This thesis aims to predict spatiotemporal LTE DL user throughput in a selected case study area in the city of Addis Ababa. The area is 1km × 1km and is located in downtown Addis Ababa. This area was chosen because it contains active business areas and the sites that serve the area have high data traffic. In addition, the proposed model is built using one specific DNN algorithm, i.e., ConvLSTM.

1.6.2 Limitation

Since drive-test is used as one means of getting data, there is an inevitable geographic limitation. Although walk tests were carried out to collect samples from areas inaccessible to vehicles, measurements from indoor environments remain unaccounted for. On top of that, spatial mapping of user throughput is based on the assumption of uniform user distribution. Furthermore, the study mainly relies on DL user throughput data obtained from NMS and does not take into account other factors that affect user throughput, such as the active number of users and radio-related metrics like channel quality indicator.

1.7 Contribution

The results of this work could be used to generate average user throughput experience in a low-resolution spatial grid. This enables operators to monitor QoE, efficiently use resources, and precisely plan future networks. Since the thesis is conducted in a selected area in Addis Ababa, the model developed would apply to this extent. However, by gathering and processing the required data, the proposed approach could be implemented anywhere.

1.8 Thesis Outline

The rest of the thesis is structured as follows. In the following chapter, the theoretical background on LTE networks, and a brief description of integrity KPIs and their formulation are discussed. Chapter three gives a summary of throughput and traffic forecasting techniques. In Chapter four, a detailed description of the followed-up approach, the model development procedure, and the obtained results are presented. Chapter five highlights the drawn conclusions and future research directions.

2 Background on LTE Network

2.1 LTE Network Overview

LTE is the standardization process for mobile communication systems by 3GPP [10]. 3GPP Release 8 standardizes the first LTE specifications, paving the way for fourth-generation (4G) technologies. LTE has evolved from its predecessors, such as Wideband Code Division Multiple Access (WCDMA) and High-Speed Packet Access (HSPA). It uses an all-Internet Protocol (IP) network architecture rather than the circuit switching scheme of its predecessors to achieve high data rates and low latency. LTE with its System Architecture Evolution (SAE), encompasses the core network element Evolved Packet Core (EPC) and the radio network, both of which are packet-switched [10].

2.1.1 Evolution towards LTE and Enabling Technologies

New technological solutions in the radio interface and access networks, such as orthogonal frequency division multiplexing (OFDM), multiple-input multiple-output (MIMO), high rate modulation, and others, enable LTE to perform better. These main technological advancements are briefly explained below [11].

- **Multicarrier air interface** – the radio access network in LTE employs OFDMA-based modulation and multiple access schemes for the DL and Single Carrier Frequency Division Multiple Access (SC-FDMA) for the uplink (UL). To achieve a high data rate low-cost and power-efficient hardware, as well as OFDM-based technology are used. In OFDMA, the 20 MHz channel is subdivided into many sub-channels and is assigned to different users. These sub-channels are used to send data simultaneously, resulting in a high data rate.

- **MIMO with 64-state quadrature amplitude modulation (64 QAM)** – The combination of 64 QAM and MIMO results in a theoretical rate of 42 Mbps, that is, 2 × 21.6 Mbps.
- **Dual cell operation** – Dual Cell High-Speed DL Packet Access (DC-HSDPA) is a feature that is further enhanced in Releases 9 and 10. It allows a UE to efficiently use two 5 MHz carriers. Assuming that both are using 64 QAM (21.6 Mbps), the theoretical data rate is 42 Mbps.
- **Further power and battery enhancements** – it employs features known as enhanced fast dormancy as well as enhanced Radio Resource Control (RRC) state transitions.

Table 2.1 summarizes LTE key features.

Table 2.1 Key features of LTE [10]

Access scheme	UL	SC-FDMA
	DL	OFDM
Bandwidth	1.4, 3, 5, 10, 15, 20 MHz	
Sub-carrier spacing	15 KHz	
Cyclic prefix length	Short: 4.7 μsec, long: 16.7 μsec.	
Modulation	Quadrature Phase Shift Keying (QPSK), 16QAM, 64QAM	

The features that contribute to the HSPA peak data rate evolution from 3GPP Release 5 up to Release 8 are shown in Figure 2.1.

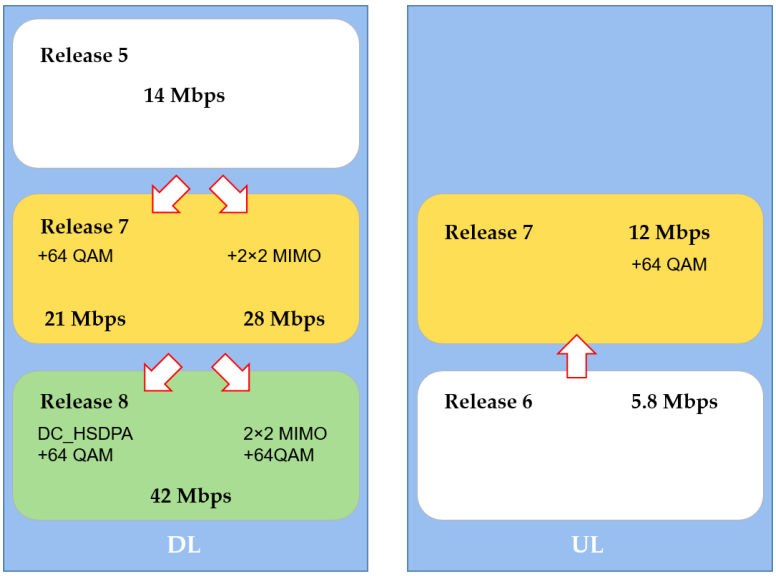

Figure 2.1 LTE evolution and enabling technologies [12]

2.1.2 LTE Network Architecture

The LTE network is based upon the Evolved Packet System (EPS) [13]. The EPS consists of the radio access network known as the Evolved Universal Terrestrial Radio Access Network (E-UTRAN), and an IP core network: the EPC. The overall network architecture is shown in Figure 2.2.

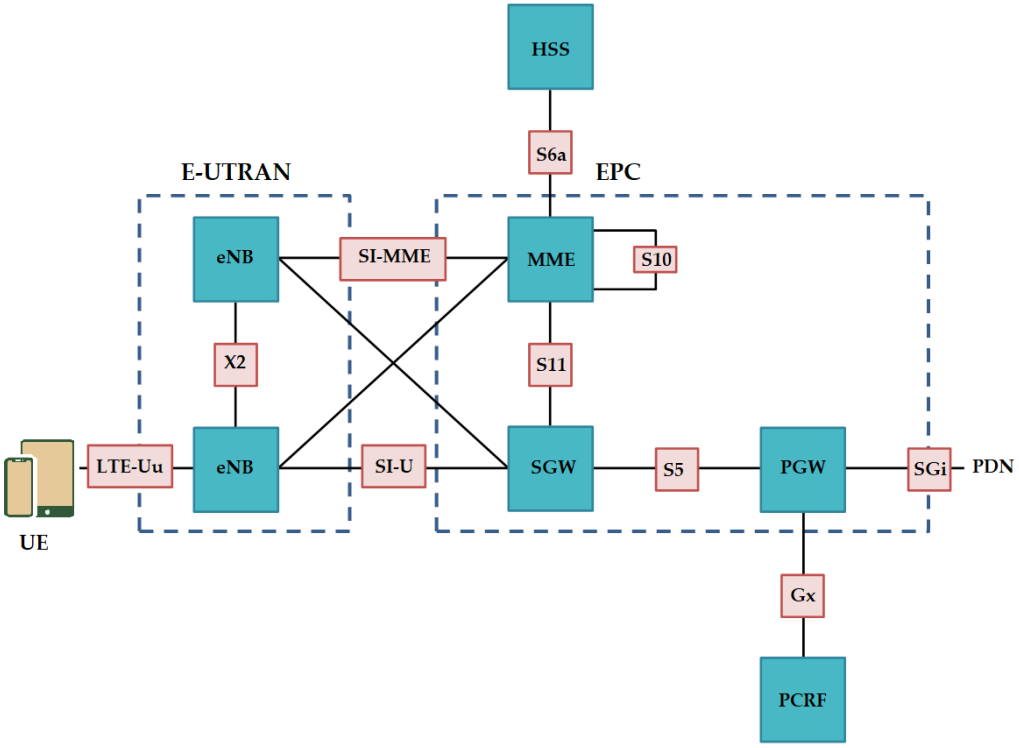

Figure 2.2 LTE network architecture [13]

A. The Core Network

The EPC consists of several functional units that are responsible for the overall control of the UE and the establishment of bearers [13].

- **The Serving Gateway (S-GW)** – is the node in the user plane that connects the EPC to the E-UTRAN. The S-GW forwards all users IP packets and acts as a mobility anchor when UEs move between evolved NodeBs (eNodeBs). It also has the information about the bearers when the UE is in an idle state and temporarily buffers DL data while the Mobility Management Entity (MME) is paging the UE to re-establish the bearers.
- **The MME** – is the EPC's control-plane node. It is in charge of bearer management, connection management, and security keys. The protocols running between the UE and the core network are called the Non-Access Stratum (NAS) protocols.

- **The Packet Data Network Gateway (P-GW)** – connects the UE to external Packet Data Networks (PDNs) and serves as the UE's default Internet router. The PGW is also in charge of allocating IP addresses to mobile devices.
- **The Policy Control and Charging Rules Function (PCRF)** – is a software component that efficiently accesses subscriber database and other specialized functions for policy control decision making.
- **The Home Subscriber Service (HSS)** – is a database that stores subscriber information. It is also responsible for generating security information and performing user authentication.

B. Radio Access Network

The LTE radio access network has a flat architecture and is consists of a network of eNodeBs. LTE integrates the radio control function into the eNodeB. As a result, there is no separate radio controller unit [10]. The eNodeB is in charge of the following radio-related functions [10]:

1. IP header compression and encryption of the user data stream
2. Radio resource management: radio bearer control, radio admission control, connection, mobility control, dynamic allocation of resources to UEs in both UL and DL (scheduling)
3. Selection of the MME during a call
4. Transfer of paging and broadcast messages to the UEs
5. Intra-eNodeB mobility control

2.1.3 LTE Protocol Architecture

The overall protocol stack for LTE control and user planes of various interfaces are shown in Figure 2.3, and Figure 2.4, respectively.

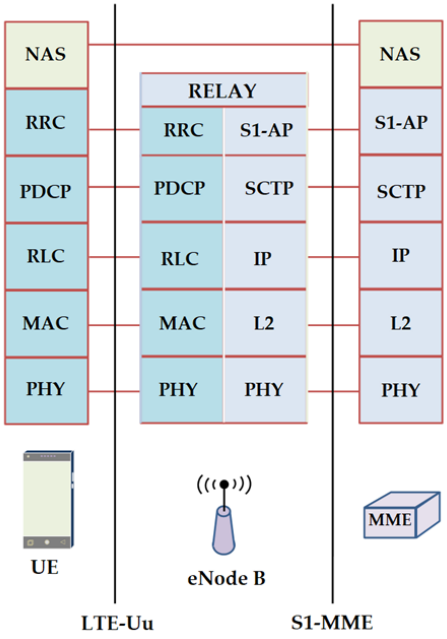

Figure 2.3 LTE control plane protocol stack [14]

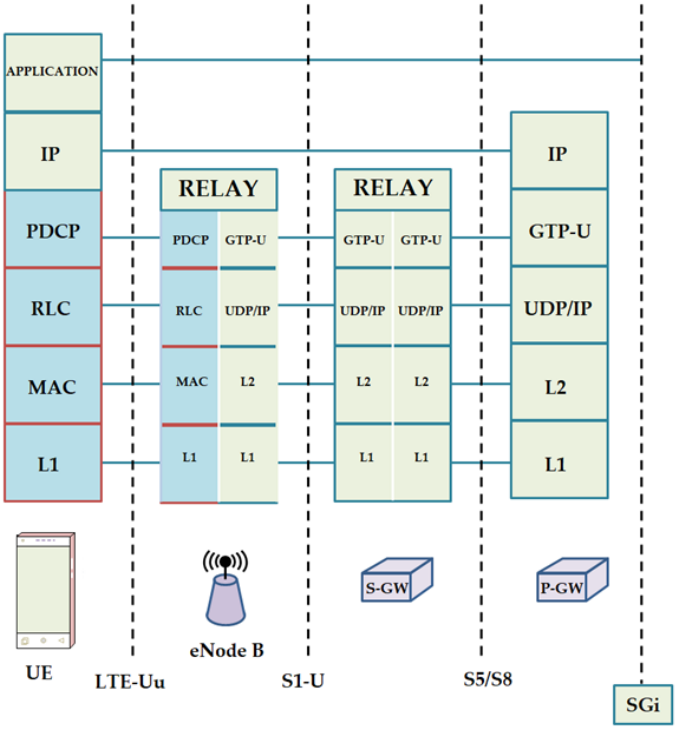

Figure 2.4 LTE user plane protocol stack [14]

- **The RRC** – is the signaling connection across the LTE-Uu interface. This sublayer's primary services and functions include mobility control, QoS management functions, radio bearer control, connection management, and measurement control.
- **Packet Data Convergence Protocol (PDCP)** – performs functions such as robust packet header compression and decompression, ciphering, deciphering, and transfer of user data.
- **The Radio Link Control (RLC)** – performs segmentation and concatenation to optimize the use of the available resource and tracks packets that were sent or received.
- **The MAC (Medium Access Control) sub-layer** – is in charge of UL and DL scheduling, as well as logical channel multiplexing and Hybrid Automatic Repeat Request (ARQ) retransmissions. It also decides which transport channel to use.

- **The function of the PHY (Physical) sub-layer** – is to provide data transport services on physical channels to the upper RLC and MAC sub-layers.
- **The NAS** – lies between UE and MME. It performs functions independent of the access technology. These functions include authentication, EPS bearer management, security control, idle mode paging origination.
- **The S1-AP (S1 Application Protocol)** – processes individual connections and then multiplexes multiple individual connections to the Stream Control Transmission Protocol (SCTP).
- **The GTP-U (GPRS Tunneling Protocol-User) protocol** – has an inherent facility to identify tunnels and also to facilitate intra-3GPP mobility. The user plane uses GTP-U over User Datagram Protocol/IP (UDP/IP) for channeling user subscriber data over the S1 interface.

2.1.4 LTE Frame Structure

There are two types of radio frame structures in LTE for physical layer transmission [11]. The type 1 radio frame structure is used for FDD (frequency division duplex), and the type 2 radio frame structure is used for LTE-TDD (time division duplex). Type 1 radio frame is 10ms in length and consists of 20 slots, each lasting 0.5 ms. Two adjacent slots form a sub frame with a length of 1 ms. Each slot contains a certain number of OFDM symbols according to the cyclic prefix length. LTE defines two cyclic prefix lengths, the normal cyclic prefix, and the extended cyclic prefix, corresponding to seven and six OFDM symbols per slot respectively [13].

For FDD operation, 10 sub frames are available for DL transmission and 10 sub frames are available for UL transmission, each separated in the frequency domain. Type 1 radio frame structure is depicted in Figure 2.4.

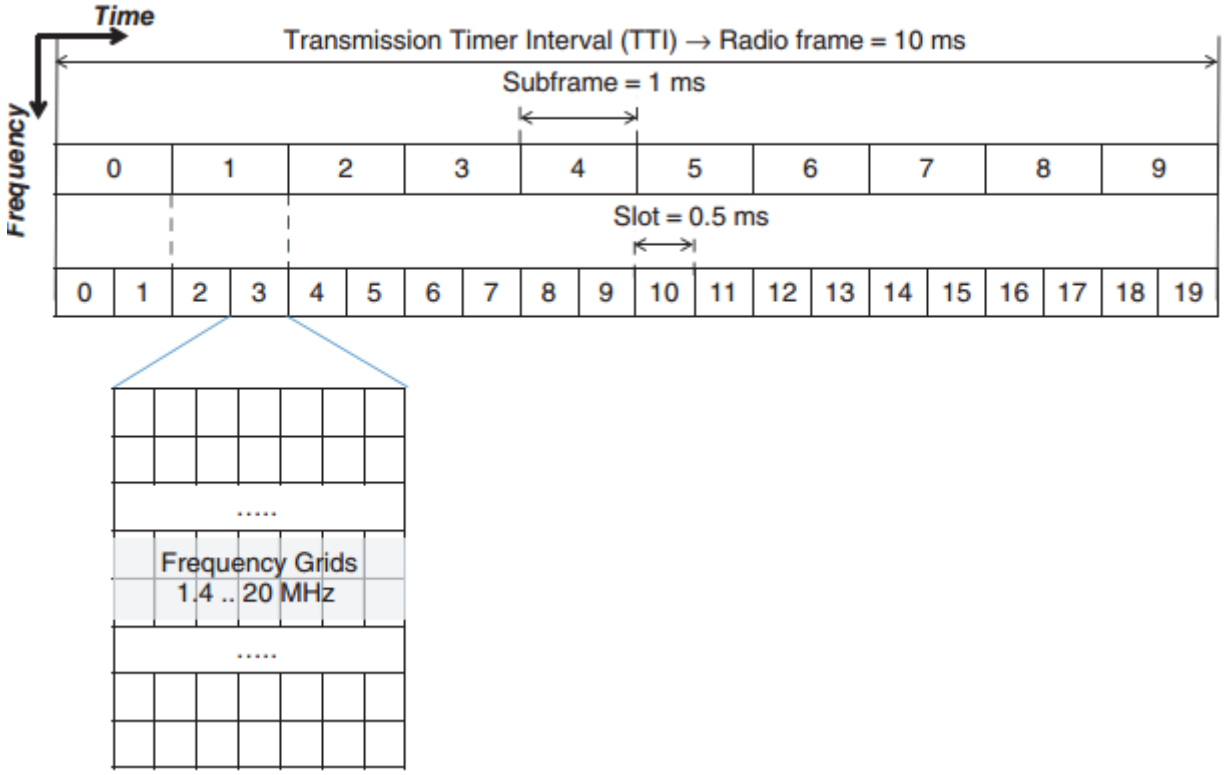

Figure 2.5 LTE FDD frame structure [11]

2.2 LTE Network Performance Measurement

A key process in cellular network operation is the assessment of any newly added network feature via service and operation KPIs [15]. These KPIs are gathered across the entire network using performance counters, considering all types of traffic. The network segments whereby KPIs could be collected include UMTS Radio Access Network (UTRAN), GSM/EDGE Radio Access Network (GERAN), Circuit Switched (CS) core, Packet Switched (PS) core, IP Multimedia Subsystem (IMS), and E-UTRAN [15]. There are five types of service performance indicators: availability, accessibility, retainability, mobility, and integrity [2]. Integrity refers to the degree to which a service is provided without excessive impairments, once obtained. Throughput falls under integrity KPI category.

2.2.1 Throughput Performance Measurement

Average user throughput is currently used by operators to verify the integrity of data services in all radio access technologies, while, *average cell throughput* is used to estimate the maximum cell capacity for network dimensioning [15].

The eNodeB allocates proportions of the available cell resources to the UEs within that cell. The more resource a UE is allocated, the more data it can send or receive per time step [8]. The allocation of radio resources is mainly affected by:

- The characteristics of the cell the UE is connected to
- The UE position relative to the serving cell and the radio environment
- The number of UEs connected to the same cell and their activities
- The modulation and coding scheme of the transmitted data

The remainder of this section discusses the indicators used in throughput analysis, which are generated from Performance Management (PM) counters in the NMS. All the considered throughput performance indicators are defined in [2] and are implemented by most vendors. The following definitions are valid for DL and UL channels [15].

A. Average User Throughput

Average user throughput is measured as:

$$AvgUeThp = \frac{TotPDCPVolDataExc}{TotEffectiveTime}[Kbps] \quad \textbf{(2.1)}$$

Where $TotPDCPVolDataExc$ is the total PDCP Service Data Unit (SDU) volume transferred per cell and Reporting Output Period (ROP), excluding data transferred in the TTI emptying the buffer (also named 'last TTIs'), and $TotEffectiveTime$ is the time used to send the information excluding last TTIs. Both $TotPDCPVolDataExc$ and $TotEffectiveTime$ are aggregated measurements of all users in a cell. Last TTIs are excluded to remove TTIs that are not fully utilized to achieve a throughput measure independent of file size [2].

B. Average Cell Throughput

Average cell throughput is calculated as:

$$AvgCellThp = \frac{TotPDCPVolData}{ActiveSchdTime}[Kbps] \quad \textbf{(2.2)}$$

Where $TotPDCPVolData$ is the total PDCP SDU data volume transferred in a cell and ROP considering all TTIs and $ActiveSchdTime$is the scheduler activity time. $ActiveSchdTime$ is incremented by 1 ms in every TTI with data to be scheduled. Therefore, $AvgCellThp$ is the aggregated user throughput in a cell when resources are being scheduled.

C. Average Radio Link Throughput

Average radio link throughput indicates the spectral efficiency in a cell and is defined as:

$$AvgRadioThp = \frac{TotSuccVolData}{TotResourcesForTr}\left[\frac{kb}{RE}\right] \quad \textbf{(2.3)}$$

Where $TotSuccVolData$ is the total successfully transmitted MAC Protocol Data Unit (PDU) data volume, and $TotResourcesForTr$ is the number of Resource Elements (REs) used for transmission. RE is the smallest assigned resource unit, consisting of one subcarrier (15 kHz) for a duration of one OFDM symbol [13]. $TotSuccVolData$ is incremented when a transmission is acknowledged on HARQ level and $TotResourcesForTr$ is incremented when HARQ feedback for a transmission is received, regardless of the feedback received. Thus, $AvgRadioThp$ indicates the average data volume transmitted per RE, as a measure of spectral efficiency.

D. Traffic

Traffic is the total user data volume in a period, and is measured as:

$$Traffic = \frac{TotPDCPVolData}{PeriodDuration}[Kbps] \quad \textbf{(2.4)}$$

Where $TotPDCPVolData$ is the transferred PDCP SDU data volume and $PeriodDuration$ is the measurement period. $TotPDCPVolData$ is the same statistic used in $AvgCellThp$ and $PeriodDuration$ is the ROP. $Traffic$ indicates the average cell throughput including periods where the scheduler is inactive.

E. Average Number of Users

Two indicators are used to measure the average number of users in one cell, namely:

$$AvgNoUeperTTIinROP = \frac{TotNoActiveUsers}{PeriodDuration} \quad \textbf{(2.5)}$$

$$AvgNoUeperTTIinSchedAct = \frac{TotNoActiveUsers}{ActiveSchedTime} \quad \textbf{(2.6)}$$

Where $TotNoActiveUsers$ is the sum of the number of active users per TTI across the whole period, $PeriodDuration$, and $ActiveSchedTime$is the scheduler activity time during this period. The main difference between them is that Equation (2.6) only takes into account TTIs with active users.

In this chapter, a comprehensive analysis of throughput network performance and how they are measured is mentioned. This allows us to figure out the difference among the different metrics, and select the right KPI for the desired objective.

Average user throughput ($AvgUeThp$) for DL is selected for the model in this thesis since a substantial amount of measurement data for the KPI can be obtained in MNO's NMS. This data can be leveraged to effectively train the proposed DNN.

In the next chapter, the characteristics of user throughput, and traffic as well as the different modeling approaches will be discussed.

3 Throughput and Traffic Prediciton Techniques

3.1 Mobile Data Throughput and Traffic Characteristics

Knowing the traffic characteristics at each site or geographic area and at different times of a day aids in modeling traffic demand in both spatial and temporal domains. This enables Internet Service Providers to develop strategies to ensure a great customer experience [16].

The aggregated mobile traffic of sites reveals a fundamental temporal pattern that is closely related to human sleep patterns. Furthermore, aggregated network traffic over an X × Y grid reveals a statistical dependence between proximate cells, whereas traffic patterns at distant cells provide little insight into how traffic consumption will evolve at a 'target' cell [16]. Similarly, research conducted in Addis Ababa, Ethiopia, shows that mobile data traffic follows the same usage pattern every day of the week, reaching its peak around 10:00 PM and its lowest value in the morning around 4:00 AM [9]. Exploratory analysis of LTE average DL user throughput in the selected target area also showed a daily seasonality. The highest throughput values were recorded after midnight (00:00 AM - 03:00 AM), inherently due to light network load. However, the afternoon (3:00 PM-6:00 PM) had the lowest throughput values.

The remaining sections of this chapter highlight the various time series, spatial, and spatiotemporal forecasting techniques that apply to mobile traffic and throughput.

3.2 Time Series Forecasting

A time series is defined as a sequence of values, arranged chronologically, and observed over time [17]. In time-series data, the behavior of past events can be used for forecasting future events since it is expected that, in the future, the influence of the behavior of past

events will still occur [18]. These behaviors are linked to time-series data components, which include: *trend, cyclical, seasonal, and irregular* components [19].

- **Trend** - a trend is a long-term propensity, increase and decrease in the time-series data. Trend component represents the structural variations of low-frequency time series.
- **Cyclical** - cyclic patterns indicate medium-term fluctuation. The cyclical pattern display increase and falls without a specified period.
- **Seasonal** - a seasonal component is a variation in time-series data that is influenced by seasonal factors such as year, quarter, month, week, day, hour. The seasonal component has stabled variation at the seasonal interval within the time series data.
- **Irregular** - an irregular component is residual time series after removing trend and seasonal components.

Time-series data forecasting is a part of statistical modeling that is widely used in various fields because of its benefits in decision-making [18]. Time series prediction models for network traffic can be categorized under four broad categories; namely, linear time series models, nonlinear time series models, hybrid models and decompose models [19]. Each of them are briefly explained in the below sections.

3.2.1 Linear Time Series Models

In linear time series prediction models, prediction of the future values is constrained to be linear functions of past observations. Traditional statistical models such as moving average, exponential smoothing, and ARIMA are among these techniques [19].

3.2.2 Nonlinear Time Series Models

Nonlinear time series are formulated by non-linear dynamic equations. To overcome the limitations of the linear models, several classes of nonlinear models have been proposed in the literature. These include the bilinear model, the threshold autoregressive (TAR), and most recently artificial neural networks (ANNs) [20].

3.2.3 Hybrid Time Series Models

A hybrid model combines two or more models. The hybrid model exploits the unique feature and strength of linear and nonlinear models in determining different patterns. Thus, it could be advantageous to model linear and nonlinear patterns separately by using different models and then combine the forecasts to improve the overall modeling and forecasting performance [20].

3.2.4 Decomposed Time Series Models

In a decomposed model, the time series data is decomposed into different components like the trend, seasonal, and cyclical. To obtain a prediction, the future values are calculated for each component and added up again [19].

3.3 Spatial Prediction

Spatial analysis is a study that provides a unique set of techniques and methods for analyzing events that are located in geographical space [21]. Cellular network traffic fluctuations show an obvious different characteristic and exhibit a strong non-uniform pattern over the space [22]. The widely used techniques for spatial modeling of mobile traffic as described in [22] are a stochastic geometry approach; namely, Spatial Poisson Point Process (SPPP), and statistical distribution approaches; namely, Lognorm distribution model, Exponential distribution model, and Gaussian distribution model.

3.4 Spatiotemporal Prediction

The spatiotemporal data traffic prediction approach incorporates both spatial correlation and temporal dependencies observed in mobile data traffic to predict the traffic demand at a given time and space [9].

However, given the complex spatiotemporal variability of mobile traffic due to user mobility and diverse network demand by different applications, the prediction process is very difficult. However, machine learning and computer vision advancements have resulted in astounding results in image recognition, natural language processing, and mobile traffic analytics [6]. Hence, machine learning techniques can be applied to solve spatiotemporal throughput prediction problems.

According to [8] existing throughput prediction algorithms can be classified into three categories: formula-based, history-based (statistical), and machine learning-based.

In a formula-based approach, a mathematical model to estimate achievable throughput is developed based on knowledge and assumption of underlying protocols. In a history-based approach, throughput is predicted from past measurements using methods like ARIMA and Seasonal Autoregressive Integrated Moving Average (SARIMA). The machine learning-based approach is a state-of-the-art and employs machine learning algorithms like RF, SVR, and various DNN algorithms.

3.5 Deep Learning and Mobile Networks

3.5.1 Overview of Deep Learning

AI, machine learning and deep learning are all terms that are frequently mentioned when discussing big data, data analytics and advanced technology [23]. The high-level relationship between the three technologies is illustrated in Figure 3.1.

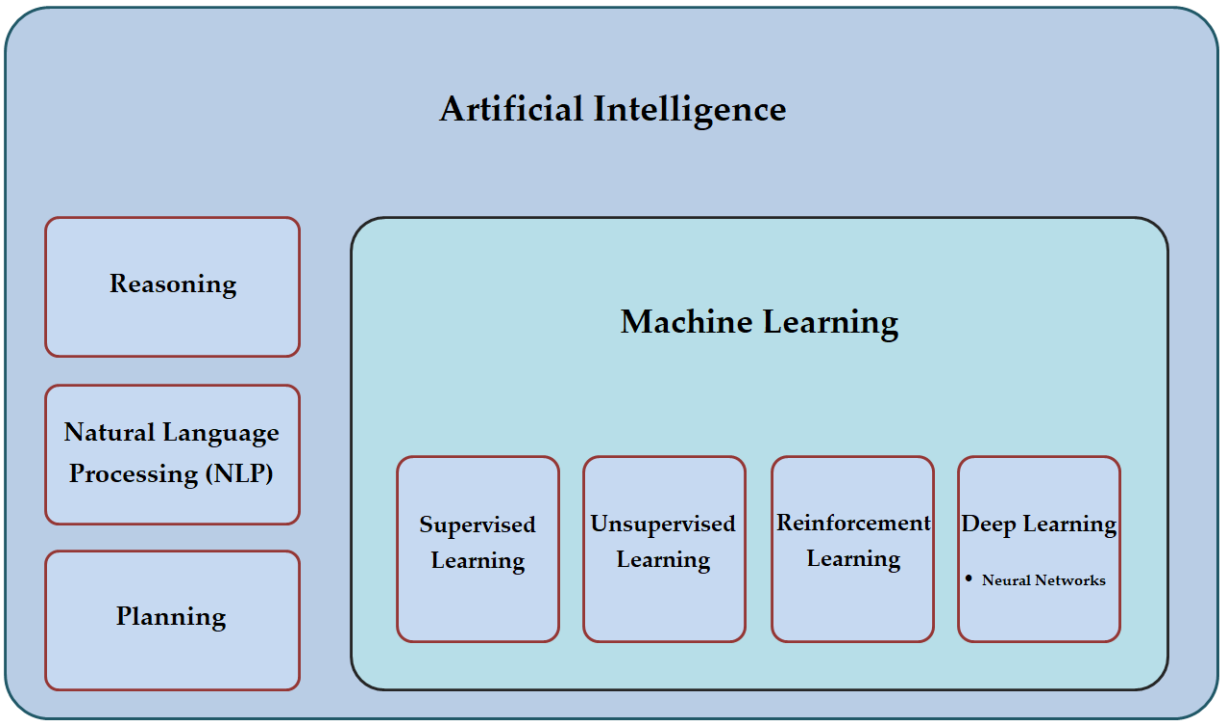

Figure 3.1 Relationship between AI, machine learning, and deep learning [23]

Deep learning is a sub-branch of machine learning that enables an algorithm to make predictions, classifications, or decisions based on data. In contrast to traditional machine learning tools, deep learning algorithms hierarchically extract knowledge from raw data through multiple layers of nonlinear processing units [23]. The most well-known deep learning models are neural networks (NNs), but only NNs that have a sufficient number of hidden layers (usually more than one) can be regarded as 'deep' models [5]. A neural network consists of three or more layers: an input layer, one or many hidden layers and an output layer. Data is ingested through the input layer. Then the data is modified in the hidden layer and the output layers based on the weights applied to these nodes [23]. Typical neural network architecture is depicted in Figure 3.2.

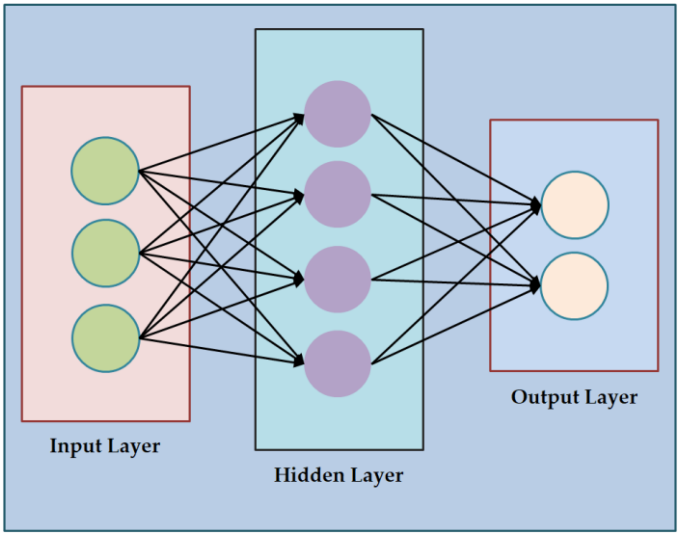

Figure 3.2 The architecture of a neural network [23]

Developing a machine learning application or putting a machine learning algorithm into action is an iterative process [23]. We can't just train a model once and forget about it; data changes, preferences shift and new competitors emerge. As a result, once a model goes into production, we must keep it fresh [23]. The steps in the machine learning cycle are illustrated in Figure 3.3.

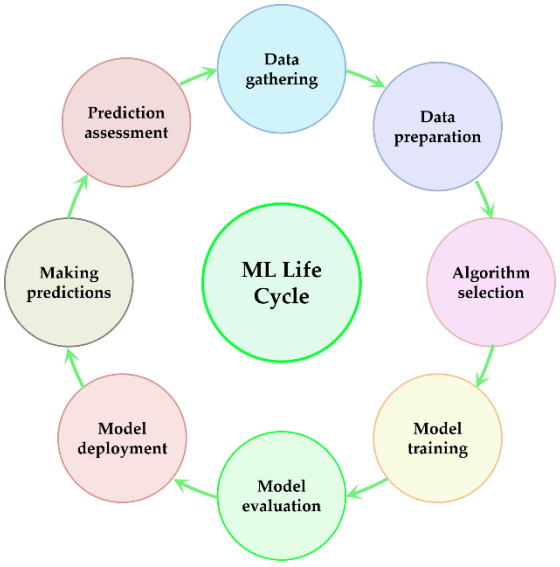

Figure 3.3 Machine learning life cycle [23]

3.5.2 Deep Learning in Mobile Networks

Deep learning can be used in a variety of mobile and wireless networks. Deep learning-driven analysis of network-level mobile data, app-level mobile data, user localization, wireless sensor networks, network control, network security and emerging mobile network applications are among them [5].

Network-level mobile data encompasses logs generated by Internet service providers, such as infrastructure metadata, network performance indicators and call detail records (CDRs) [5]. Deep learning's recent remarkable success has sparked global interest in exploiting this methodology for mobile network-level data analysis and optimizing mobile network configurations, thereby improving end-user QoE. Experiments with a real-world dataset show that it outperforms Support Vector Machine (SVM) and ARIMA. Deep learning was also employed in [6] [24], where the authors employ CNNs and LSTMs to perform mobile traffic forecasting. By effectively extracting spatiotemporal features, their proposals gain significantly higher accuracy than ARIMA. A key finding is that large-scale mobile network traffic can be processed as sequential snapshots, which resemble images and videos. These correlations can be effectively learned by CNNs and RNNs, as they are specialized in modeling spatial and temporal data (e.g., images, traffic series). The deep learning algorithms applicable for network-level spatiotemporal prediction are discussed in the following section.

A. Convolutional Neural Network

CNNs are a type of neural network that was designed to efficiently handle image data. They have proven effective in challenging computer vision problems both achieving state-of-the-art results on tasks like image classification [25]. The structure of a basic CNN as shown in Figure 3.4 comprises convolutional, pooling, and fully connected layers besides input and output layers.

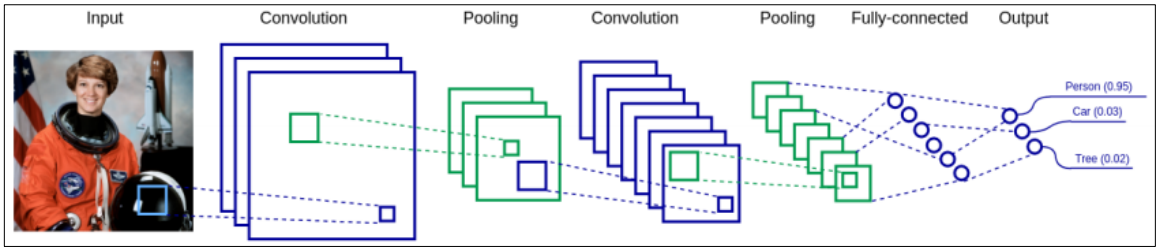

Figure 3.4 Structure of basic CNN [26]

- **Convolutional layer** – the convolutional layer is the most important building block of a CNN. It consists of a set of filters (kernels), where each filter is applied across all areas of the input data to capture correlations between different data regions. Mathematically, for each location p_y of the output y, the convolution performs the following operation:

$$y(p_y) = \sum_{p_G \in G} w(p_G) \cdot x(p_y + p_G) \tag{3.1}$$

where p_G denotes all positions in the receptive field G of the convolutional filter w, effectively representing the receptive range of each neuron to inputs in a convolutional layer. The weights w are shared across different locations of the input map. The operation of a 2D convolutional layer is depicted in Figure 3.5. The inputs of a 2D CNN layer are multiple 2D matrices with different channels, e.g., RGB value of images. A convolutional layer applies multiple filters shared across different locations, to "scan" the inputs and produce output maps.

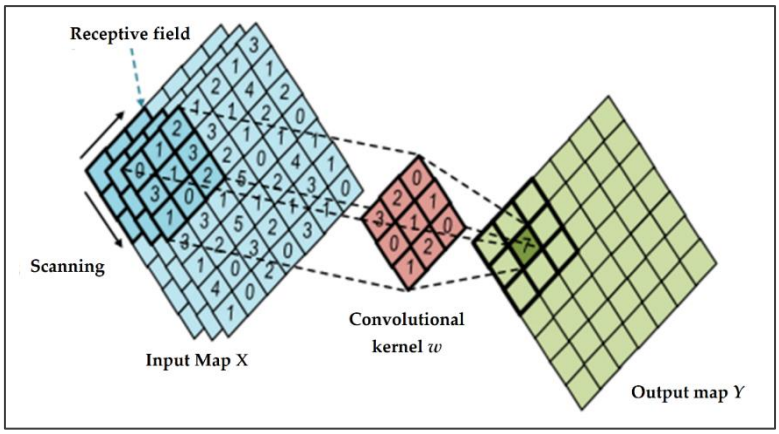

Figure 3.5 Operational principle of a 2D convolutional layer [5]

- **Pooling layer** - a limitation of the feature map output of convolutional layers is that they record the precise position of features in the input. This means that small movements in the position of the feature in the input image will result in a different feature map. This can happen with re-cropping, rotation, shifting, and other minor changes to the input image. This problem can be addressed by adding a pooling layer after the convolutional layer [27]. A pooling layer generates a separate down sampled and lower resolution representation of convolutional layer output. The pooling operation is based on down sampling dimension (stride), and pooling type: maximum pooling or average pooling.
- **Fully connected layer** - a fully connected layer is a layer that is at the end of CNN and interprets the features extracted by the layer that precedes it [25].

B. Recurrent Neural Network

Recurrent Neural Networks (RNNs) are designed for modeling sequential data, where sequential correlations exist between samples. At each time step, they produce output via recurrent connections between hidden units, as shown in Figure 3.6.

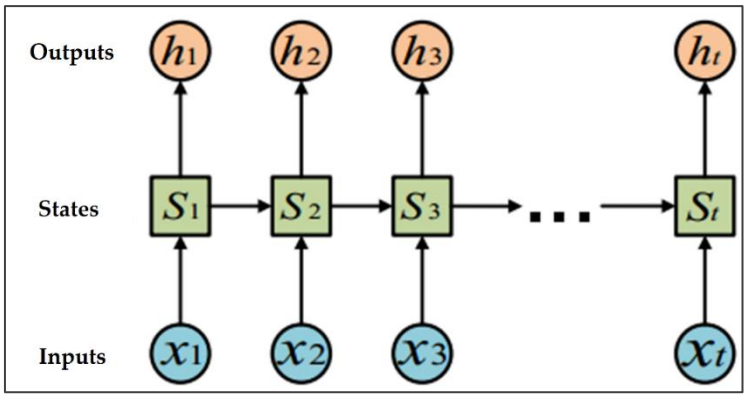

Figure 3.6 Structure of a simple RNN [5]

Given a sequence of inputs $x = \{x_1, x_2, \ldots, x_T\}$, a standard RNN performs the following operations:

$$S_t = \sigma_s(W_x x_t + W_s s_{t-1} + b_s)$$
$$h_t = \sigma_h(W_h s_t + b_h) \quad \textbf{(3.2)}$$

Where S_t represents the state of the network at time t and it constructs a memory unit for the network. Its values are computed by a function of the input x_t and the previous state s_{t-1}. h_t is the output of the network at time t. The weights W_x, W_h and biases b_s, b_h are shared across different temporal locations. This reduces the model complexity and the degree of overfitting. The RNN is trained via a backpropagation through Time (BPTT) algorithm. However, gradient vanishing and exploding problems are frequently reported in traditional RNNs.

C. LSTM

LSTM is a special kind of RNN with a set of gates to mitigate vanishing and exploding gradient problems of RNN. This is because LSTMs can handle long-term dependencies due to a specially designed memory cell. A standard LSTM performs the following operations:

$$i_t = \sigma(W_{xi} X_t + W_{hi} H_{t-1} + b_i)$$
$$f_t = \sigma(W_{xf} X_t + W_{hf} H_{t-1} + b_f) \quad \textbf{(3.3)}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot tanh(W_{xc}X_t + W_{hc}H_{t-1} + b_c)$$
$$o_t = \sigma(W_{xo}X_t + W_{ho}H_{t-1} + b_o)$$
$$h_t = o_t \odot tanh(C_t)$$

Where $\odot$ denotes the Hadamard product, C_t denotes the cell outputs, h_t are the hidden states, i_t, f_t, and o_t are input gates, forget gates, and output gates, respectively. $\sigma(\,)$ is the standard logistic sigmoid function, and $tanh(\,)$ is the hyperbolic tangent function. Parameters $W_j(j = xf, hf, xi, hi, ox, oh)$ denote the weight matrices in each layer and $b_j(j = f, i, o)$ denote the bias of each gate. The forget gate decides whether we want to erase the cell state or not, the input gate decides what new information is going to be added to the memory cell, and the output gate decides what the total cell output is going to be. The internal structure of LSTM cell is illustrated in Figure 3.7.

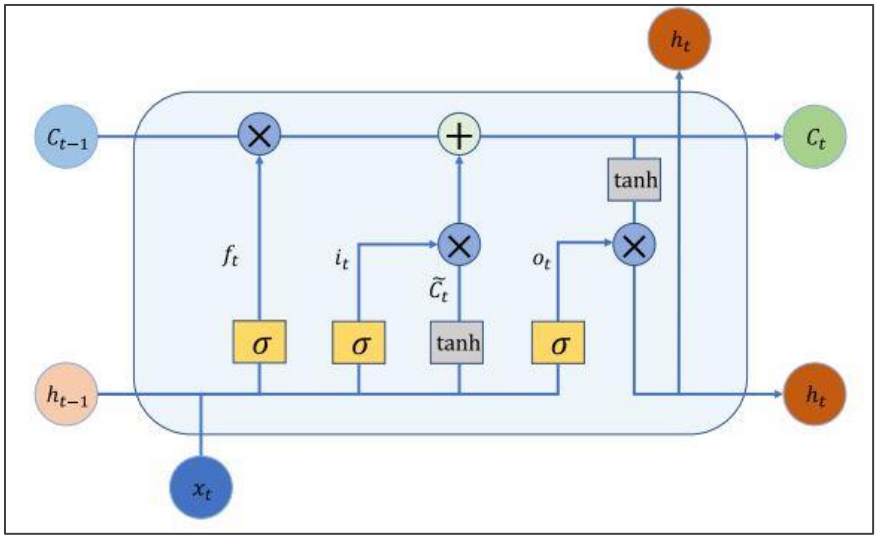

Figure 3.7 The inner structure of LSTM cell [28]

D. ConvLSTM

For spatiotemporal problems, where a temporal point in the input sequence is twodimensional, LSTM models face similar problems as feed-forward neural networks [29]. with spatial input: they cannot efficiently capture spatial correlations and generally have too many connections. To overcome these issues, ConvLSTM structure was designed. It maintains the same structure as LSTM but replaces linear

weight operations with convolutions. ConvLSTM can be described with the following equations, where '$*$' denotes two-dimensional convolution operation:

$$
\begin{aligned}
i_t &= \sigma(W_{xi} * \mathcal{X}_t + W_{hi} * \mathcal{H}_{t-1} + W_{ci} \odot \mathcal{C}_{t-1} + b_i) \\
f_t &= \sigma\big(W_{xf} * \mathcal{X}_t + W_{hf} * \mathcal{H}_{t-1} + W_{cf} \odot \mathcal{C}_{t-1} + b_f\big) \\
\mathcal{C}_t &= f_t \odot \mathcal{C}_{t-1} + i_t \odot tanh(W_{xc} * \mathcal{X}_t + W_{hc} * \mathcal{H}_{t-1} + b_c) \\
o_t &= \sigma(W_{xo} * \mathcal{X}_t + W_{ho} * \mathcal{H}_{t-1} + +W_{co} \odot \mathcal{C}_t + b_o) \\
\mathcal{H}_t &= o_t \odot tanh(\mathcal{C}_t)
\end{aligned}
\tag{3.4}
$$

Note that in the case of ConvLSTM, inputs $\mathcal{X}_t$, hidden state $\mathcal{H}_t$ and cell state $\mathcal{C}_t$ are three-dimensional tensors with dimensions: height × width × number of channels. It was shown that ConvLSTM outperforms LSTM models in video prediction by more successfully capturing Spatio-temporal patterns. ConvLSTM is a strong building block for models solving Spatio-temporal problems and is used throughout this work.

4 LTE Throughput Modelling and Prediction

In this chapter, the approach followed to model and predict LTE DL user throughput is discussed in detail.

4.1 Proposed Approach

The spatiotemporal modeling and prediction of LTE DL user throughput approach followed in this thesis aims to leverage big data from MNO's NMS and deep learning techniques. Temporal average user throughput data can be obtained from NMS on an hourly basis for every cell. However, the spatial mapping of this measurement is not straightforward. Several studies such as: [5], [9], [21], [24] have used the base station's geographic location to spatially aggregate site-level network traffic. This approach cannot be directly implemented for user throughput prediction since:

- A single site, in most scenarios, has more than one cell
- Each cell has distinct coverage with different azimuth and probably different radius
- Overlapping coverage regions exist at cell borders

Hence, the site's geo-location can't be merely used. Instead, an approach that takes into consideration the actual cell coverage should be followed.

The approach in this thesis is based on the above consideration. The core elements and the overall workflow of the spatiotemporal user throughput modeling and prediction process are depicted in Figure 4.1.

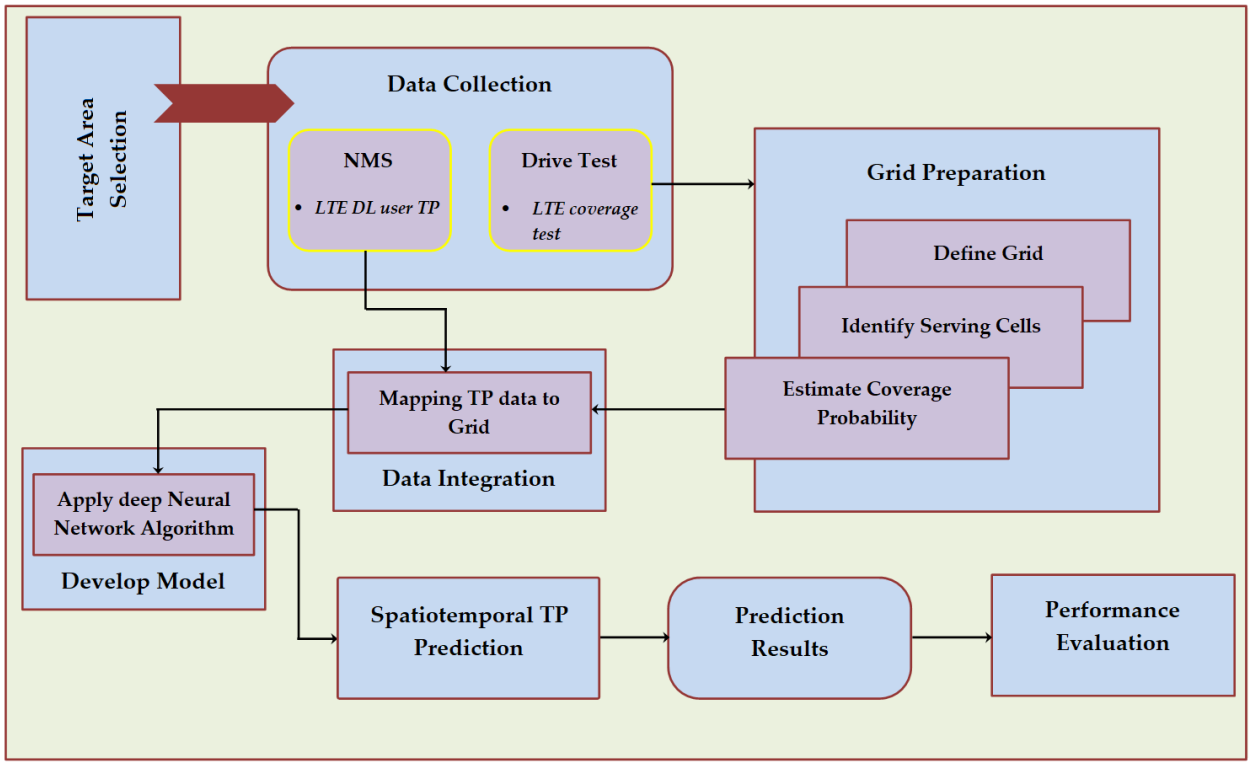

Figure 4.1 LTE DL user throughput modeling and forecasting system model

4.2 Data Collection

The first step towards data collection is identifying a target area. A 1Km by 1Km target area is selected in downtown Addis Ababa. The area is chosen based on network statistics, i.e., number of users and data traffic.

Having identified the target area, data is collected from two sources: from NMS, and drive test. Two months hourly data of LTE DL user throughput data is collected from the NMS side, and coverage-related measurements such as cell ID, RSRP, site location, and location information of every sample are collected from the drive and walk tests.

4.3 Coverage Analysis

The coverage of cellular networks is subjected to fluctuations due to unforeseen changes in the radio environment such as the construction of new buildings and interferences, or

equipment failure [30]. Consequently, cells might not cover geographic areas according to the estimation made during the planning phase, and overlapping cell coverage areas are expected. To that end, coverage analysis is necessitated to get realistic coverage capabilities of each cell serving the target area. The coverage analysis will also allow us to get coverage in a spatial grid, which is a foundation for preparing data for the proposed algorithm. Note that CNN-based algorithms such as ConvLSTM work with pixel/grid-based data.

LTE coverage test has been conducted using Nemo Handy tool within the target area. The test was conducted for two days to get more samples and ensure the integrity of radio signal measurement. Then, the target area is divided into regular grids, and samples' features, i.e., RSRP and distance falling in each grid are utilized to determine the coverage probabilities of cells in each grid. The number of null grids was compared to the grid resolution to decide the grid dimension.

The target area with pre-planned drive test routes (green paths) and walk test areas places marked with yellow pushpins are illustrated in Figure 4.2. Figure 4.3 depicts measurement samples along the chosen drive test route. The various colors of samples represent the corresponding serving cells.

The target area has been divided into 6 by 6 and 10 by 10 grids as a starting point. One null grid with a spatial grid resolution of 167m is obtained from the former approach while five null grids with a resolution of 100m are obtained from the latter. Given the higher resolution and relatively small number of null grids, a 10 by 10 grid is chosen. Increasing the number of grids has resulted in a considerable number of null grids, which leads to inaccurate modeling. Figure 4.4 and Figure 4.5 depict samples falling in 6 by 6 and 10 by 10 grids respectively.

Figure 4.2 Target area and drive test route

Figure 4.3 Drive test measurement samples

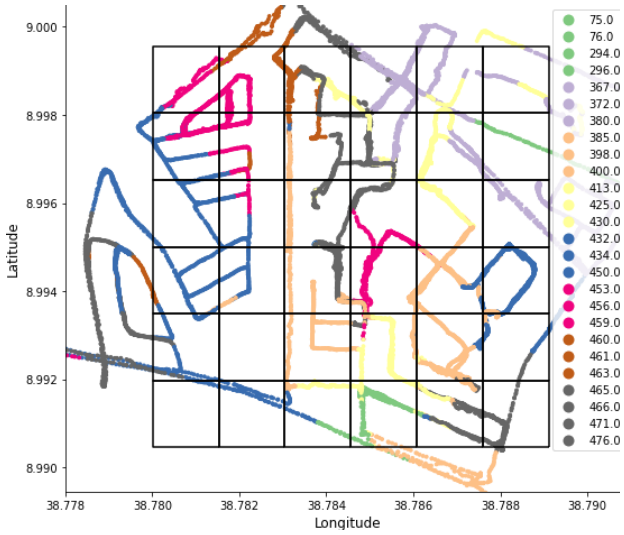

Figure 4.4 Samples and contributing cells (6 by 6 grid)

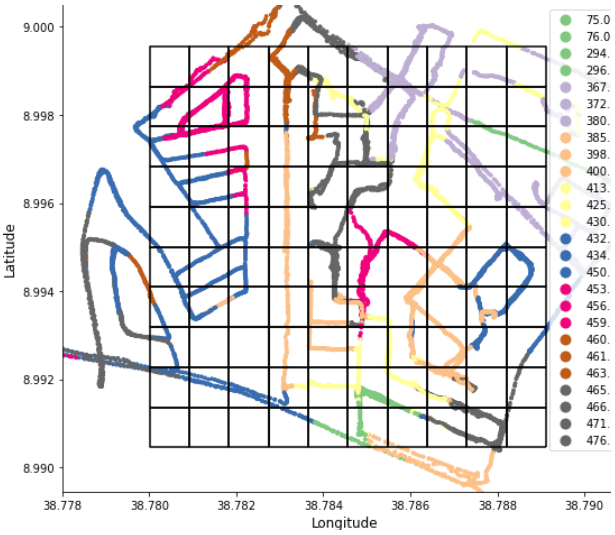

Figure 4.5 Samples and contributing cells (10 by 10 grid)

After defining geo-grids, the coverage capabilities of cells in each grid are evaluated using the approach mentioned in [31]. The method is described below:

- The latitude and longitude of the center for each grid is calculated and stored.
- The number of samples and average signal strength of each cell (i.e., S_i and R_{S_i}, $i = 1 \sim N$ number of cells serving the grid) is calculated.
- The distance between the center of the grid and each cell site (i.e., D_i , $i = 1 \sim N$) is calculated.

- The number of weighted samples of each cell is calculated as:

$$\tilde{S}_i = S_i ra_{(Rs_i)} rd_{(D_i)} \quad \textbf{(4.1)}$$

Where $ra_{(Rs_i)}$ denotes the ranking of R_{S_i} within the group arranged in ascending order, and $rd_{(D_i)}$ denotes the ranking of D_i within the group arranged in descending order.

- The coverage probability of the i^{th} cell, $CP(i)$ is then calculated as:

$$CP(i) = \frac{\tilde{S}_i}{\sum_{i=1}^{N}(\tilde{S}_i)} \quad \textbf{(4.2)}$$

Where $\tilde{S}_i$ is the number of weighted samples for each cell, and N is the number of serving cells in the grid.

4.3.1 Handling Null Grids

As shown in Figure 4.5, five grids: 0, 10, 12, 20, 22, counting from bottom to top and left to right, out of a total of 100 did not have samples. The paper in [31] dealt with the coverage of null grids based on the coverage of neighbor grids. The dominant cell in the neighbor grids surrounding the null grid was assumed to be the dominant cell. Following a similar approach, null grids are handled as illustrated in Table 4.1.

Table 4.1 Handling of null grids

Null grid	Neighbor grids	Neighbor serving physical cell identity	Null grid physical cell identity
0	1	432	432
	10	-	
	11	432	
10	0	432	432
	1	432	
	11	432	
	20	-	

	21	432	

4.4 Spatial Mapping of NMS Data

The spatiotemporal traffic distributions are governed by variations in users distribution and their respective demands, with demand referring to the actual KPI value (e.g., throughput). In most studies, probability distributions such as Poisson and Uniform are used as a convenient way of estimating user distribution. However, a more realistic user distribution is obtained by applying population data [32].

The paper in [33] formulated an approach to realistically distribute cell-level network traffic based on cell coverage, time-varying traffic usage intensity per geo-type and cell-level traffic measurement from NMS. The cell coverage has been estimated based on either dominant cell or cell assignment probabilities. Once the coverage of each cell is identified, network load-related KPIs were associated with the respective coverage area. Applicable KPIs for spatiotemporal distribution are:

- Transmitted data in total [Bit]
- Throughput in [Bit/s]
- Voice calls in [Erl]
- Voice call [attempts] in total

The distribution of KPIs is based on either assuming a uniform distribution of users within the coverage area of the cell or heterogeneous distribution of users based on land-use class, days of the week, and times of the day. For uniform distribution, the generated user traffic per cell is distributed equally for every point served by the cell. On the other hand, in heterogeneous user distribution, a fractional distribution is calculated based on the day of the week (i.e., working days, Saturdays, Sunday, and holidays), time of the day (i.e., morning, night, and rush hour), and land-use class (e.g. commercial, industrial,

building, road, etc.). The latter methodology has been implemented in [33] considering a major assumption that 80% of users in the study area reside inside any kind of building.

A method for spatial aggregation of CDR data obtained from telecommunications companies in the city of Milan was discussed in [34]. The CDR which includes Short Message Service (SMS), voice call, and internet services contains the time of interaction and serving radio base station (RBS).

To spatially aggregate CDRs, first, the coverage map of each RBS in the city is obtained. Then the geographic area of the city is divided into regular square grids. The number of CDRs in square grid i, $S_i(t)$ is then calculated as follows:

$$S_i(t) = \sum_{c \in C_{map}} R_c(t) \frac{A_{c \cap i}}{A_c} \tag{4.3}$$

Where $R_c(t)$ is the number of records in the coverage area c at time t, A_c is the surface area of coverage c, and $A_{c \cap i}$ is the spatial intersection between coverage c and grid i.

Finding a precise user distribution per clutter type, and at different times is beyond the scope of this study. Hence, uniform distribution of users is considered. Taking this assumption and adopting the above methodologies, hourly LTE average DL user throughput per grid, $DL_{AvgUeTp_g}(t)$ can be calculated as shown in Equation (4.4).

$$DL_{AvgUeTp_g}(t) = \sum_{i=1}^{N} DL_{AvgUeTp_i}(t)\, CP(i) \tag{4.4}$$

Where $DL_{AvgUeTp_i}(t)$ is hourly throughput data of cell i, $CP(i)$ is the coverage probability of cell i in grid g , and N is the number of serving cells in the grid g.

4.5 Data Preprocessing

4.5.1 Data Cleaning

The total number of samples collected via drive test is 8,171 for 'day 1' and 12,997 for 'day 2'. There were a few missing values in the features: longitude, latitude, and RSRP. Instances with missing values were simply discarded. After cleaning, the number of samples are 6498 for 'day 1' and 10,576 for 'day 2'. From the NMS side, 231 instances have null values from 112,728 records. These values are replaced by measurements obtained from the prior day with the same hour since LTE average DL user throughput has a daily seasonality.

4.5.2 Data Integration

Since the prediction is intended to be on a grid level, we need to have grid-level sequential LTE DL user throughput data as input to the DNN algorithm. The coverage probabilities of cells in each grid have been obtained using the technique discussed in Section 4.3, and for data integration, the method for mapping NMS data to grids, mentioned in Section 4.4 is implemented.

Since the data is collected for 60 days on hourly granularity, the final dataset after grid integration becomes of size: 60 × 24 × 10 × 10 = 144,100. The data integration process, as shown in Figure 4.6 is carried out based on two common features: 'PCI', and 'eNodeB ID'.

DT data after cleaning

grid_ID	PCI	eNodeB ID	CP
0	-	-	1
1	-	-	0.912
1	-	-	0.086
1	-	-	0.002
2	-	-	0.768
2	-	-	0.232

OSS data after cleaning

Time	eNodeB ID	Cell Name	DL_user_TP	PCI
4/10/2021 0:00	-	-	22.8394	-
4/10/2021 1:00	-	-	17.1316	-
4/10/2021 2:00	-	-	27.1081	-
4/10/2021 3:00	-	-	16.783	-
4/10/2021 4:00	-	-	4.7299	-
4/10/2021 5:00	-	-	22.8394	-

Time	grid_id	DL_user_TP
4/10/2021 0:00	0	20.865
4/10/2021 1:00	0	20.066
4/10/2021 2:00	1	19.255
4/10/2021 0:00	1	16.623
4/10/2021 1:00	1	18.251

Grid Throughput data

Figure 4.6 Data integration technique.

4.5.3 Data Normalization

Gradient-based training methods would require more time to tune the model weights for the outputs to match large target values [29]. Since the time-series data can have a wide range of values, it needs to be scaled/normalized to the same range of values to speed up the learning process [35]. To solve this problem, Min-Max normalization that transforms all features into the range [0,1] is used [35]. The mathematical formulation for Min-Max normalization of some feature X is depicted in Equation (4.5).

$$X_{norm} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad \textbf{(4.5)}$$

4.6 Model Development

As discussed in Section 3.5.2, DNN algorithms such as CNNs and RNNs are specialized in modeling image and time-series data. ConvLSTM, for example, is widely used in video

prediction due to its ability to capture Spatio-temporal patterns. This ability makes it preferable in spatiotemporal modeling. Several works including those in [6], [9], [21] have employed it to model and predict spatiotemporal mobile data traffic. Hence, in this study, ConvLSTM is applied to model and predict spatiotemporal LTE DL user throughput.

The model was built using open-source Python deep learning libraries Keras and Tensorflow. A network of stacked ConvLSTM algorithms is trained and evaluated to solve the spatiotemporal throughput forecasting problem. The model is optimized using various parameters to obtain the best model as described in subsequent sections.

4.6.1 Network Architecture

As illustrated in Figure 4.7, the proposed network is comprised of a stacked ConvLSTM2D algorithm in the input and hidden layers and a Conv3D algorithm in the output layer. ConvLSTM2D expects five-dimensional input data namely: *samples, time-steps, rows, columns, features* where each time-step contains a 2D (row × column) throughput data [25]. Since the output of ConvLSTM2D is a 3D tensor, Conv3D is utilized in the output layer to get a 3D output at each time step.

Training Deep Neural Networks is complicated by the fact that the distribution of each layer's inputs changes during training, as the parameters of the previous layers change. This slows down the training by requiring lower learning rates. Batch normalization enables a higher learning rate and more resilient training by normalizing activations throughout the network [36]. Hence, batch normalization is added in each hidden layer of the network.

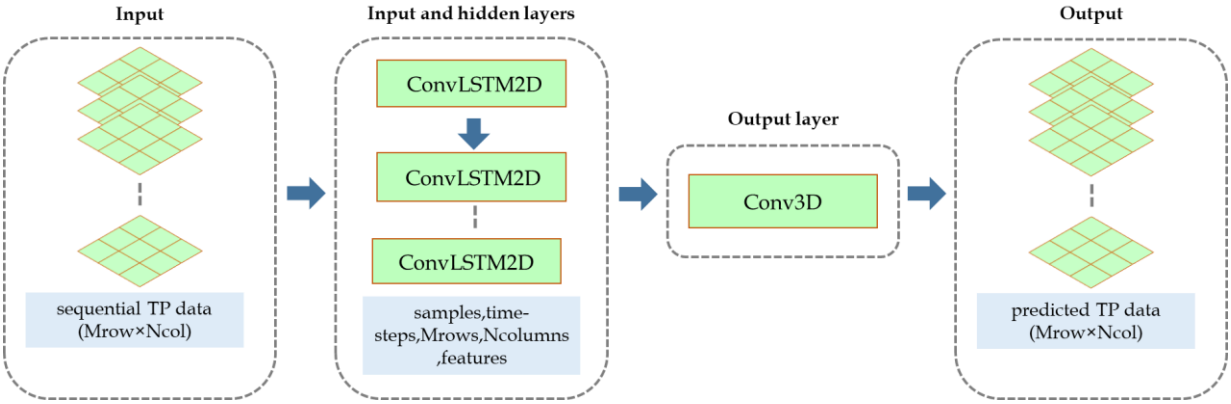

Figure 4.7 Proposed network architecture

4.6.2 Model Training

Time series data requires preparation before it can be used to train a supervised learning model, such as RNN and LSTM neural networks [25].

A supervised learning algorithm requires data to be prepared as a collection of instances, each with an input component (X) and an output component (y). The model then learns how to map inputs to outputs from training examples as shown in Equation (4.6).

$$y = f(X) \tag{4.6}$$

In this study, the data is transformed with input sequence configurations of 24 and 12-hours. After data transformation, the data is split into three sets: training (80%), validation (10%), and test (10%).

The proposed network is trained with the two input sequence configurations. While training the model with each configuration, hyper parameters of ConvLSTM algorithms were tuned with a grid search algorithm.

The training was made using Mini-Batch Stochastic Gradient Descent, with a carefully selected number of batches. Adam was chosen to update network weights since it's widely accepted in literature due to its computational efficiency [29].

The investigated hyper parameter space for ConvLSTM is shown in Table 4.2. The grid search algorithm trains the model for a list of configurations from the parameter space, evaluates the prediction results, and sorts out the best models using Mean Squared Error (MSE) evaluation metrics.

Table 4.2 Hyper parameter tuning space.

Filter	Kernel	Hidden layer	Learning rate	Batch size	Epoch	Input shape
32, 64	(3,3)	1, 2, 3, 4, 5	0.001, 0.005, 0.01, 0.1	1, 3, 5, 10, 100	50, 100, 150, 200, 300	12-hour, 24-hour

The best model, which is obtained after running the grid search algorithm has the parameters shown in Table 4.3.

Table 4.3 Best model hyper parameters

Filter	Kernel	Hidden layer	Learning rate	Batch size	Epoch	Input shape
32	(3,3)	1	0.001	3	100	12-hour

Learning curves help us to visualize whether the training and validation scores improve as we increase the number of examples [37]. This allows us to detect the overfitting and underfitting problems of a model. The learning curves for the best models are depicted in Figure 4.8 and Figure 4.9.

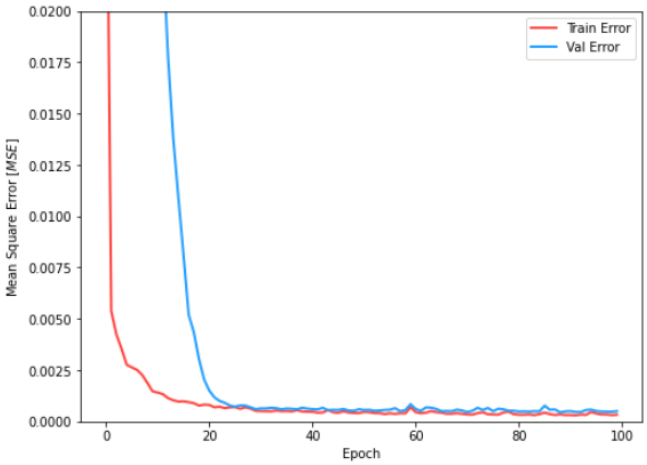

Figure 4.8 Training curve for a 12-hour input sequence

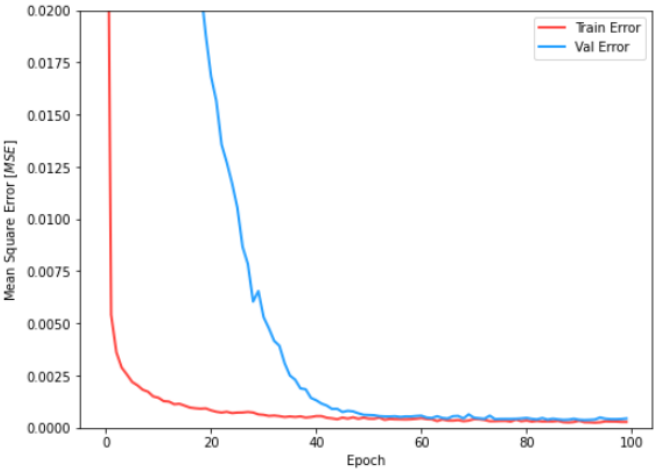

Figure 4.9 Training curve for a 24-hour input sequence

Figure 4.8 and Figure 4.9 show that not only do training and validation losses decrease, but they also become more stable as the number of training examples increases. This indicates that the trained model is a good fit model.

4.6.3 Model Performance Evaluation

A forecast problem that requires a prediction of the next time step is called a one-step forecast model. Whereas a forecast problem that requires a prediction of more than one time-step is called a multi-step forecast model [25]. The spatiotemporal throughput prediction problem in this thesis is a multi-step forecasting problem that tries to predict

the most likely *K*-steps in the future given the previous *S* observations. Since the previous error propagates to the next time step of the prediction, the performance of the model is evaluated at each time step to test the maximum capacity of the model. The time step that yields a reasonable performance is chosen as the model's prediction capacity.

There are a bunch of accuracy metrics used to evaluate the goodness of prediction of Artificial Neural Networks (ANNs). MSE, RMSE, MAE, and MAPE are the most common evaluation metrics used to evaluate ANN predictive models in engineering and other various disciplines [38]. The first three are scale-dependent metrics, whereas the last one is percentage-dependent. The formulas for the metrics are defined as [38].

$$MSE = \frac{1}{n}\sum_{1}^{n}(\hat{Y} - Y)^2 \tag{4.7}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{1}^{n}(\hat{Y} - Y)^2} \tag{4.8}$$

$$MAE = \frac{1}{n}\sum_{1}^{n}\left|\hat{Y} - Y\right| \tag{4.9}$$

$$MAPE = \frac{1}{n}\sum_{1}^{n}\left|\frac{\hat{Y} - Y}{Y}\right| \times 100\% \tag{4.10}$$

In summary, MAE gives the same weights to all errors while RMSE penalizes errors with large absolute values. Hence, MAE is better at providing average model performance. On the other hand, MAPE measures the magnitude of error in percentage terms and gives interpretable observation on the quality of prediction [38]. To that end, a combination of evaluation metrics is required to better assess model performance.

The values of MAPE and their interpretation, which are shown in Table 4.4 are used for the explanation of the prediction accuracy

Table 4.4 Interpretation of MAPE value [9]

MAPE	Interpretation
<10	Highly accurate forecasting
10 - 20	Good forecasting
20-50	Reasonable forecasting
>50	Inaccurate forecasting

The evaluation results of the trained models in this thesis are depicted in Table 4.5.

Table 4.5 Model evaluation results

Input shape	Metrics	Hour			
		1	2	3	4
24-hour	RMSE	4.11	4.65	4.53	4.56
	MAE	2.94	3.22	3.36	3.47
	MAPE	18.92	25.64	23.30	22.92
12-hour	RMSE	1.40	1.71	2.02	3.97
	MAE	1.19	1.30	1.52	2.32
	MAPE	11.68	12.63	14.68	23.20

From the evaluation results, first, it can be seen that the 12-hour input sequence model outperforms that of 24-hour input in all metrics. Next, when we look at the prediction of the 12-hour input model at different hours the MAPE value exceeds 20 at the fourth hour. This implies the forecast at this hour falls into the reasonable forecasting category (Table 4.4). Hence three hours is chosen as the prediction capability of the chosen model since we could get good forecasts.

4.7 Prediction Results

The temporal prediction results for the two models with 24-hour and 12-hour input sequences are shown in Figure 4.10 and Figure 4.11, respectively. The results are obtained from a four-hour forecast of a single cell grid. The forecast from the 24-hour input seems to perform well initially. However, the prediction diverges in the next hours and didn't capture the general trend of the original data. On the other hand, the forecast with a 12-

hour input sequence fits well with the trend of the original data. As a result, this later model is the better model, matching the findings in the performance evaluation.

To further interpret the prediction result, it's better to visualize the errors in terms of throughput in Mbps The prediction error at the third hour for the specific grid is found to be approximately 0.5 Mbps. According to [4], 1 Mbps is considered as the target speed for normal video (i.e., a resolution of 360p) to ensure smooth playback and short startup time. Hence, with the achieved prediction accuracy, it's possible to detect a service degradation in a normal video service ahead of time in the given area.

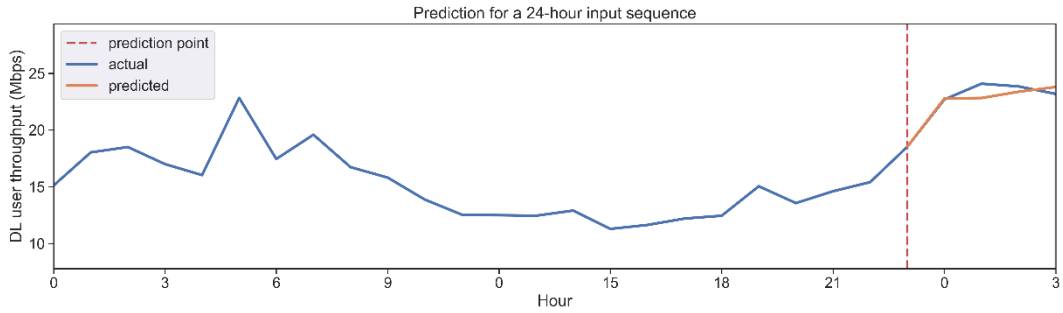

Figure 4.10 Temporal throughput prediction for one cell (24-hour observation)

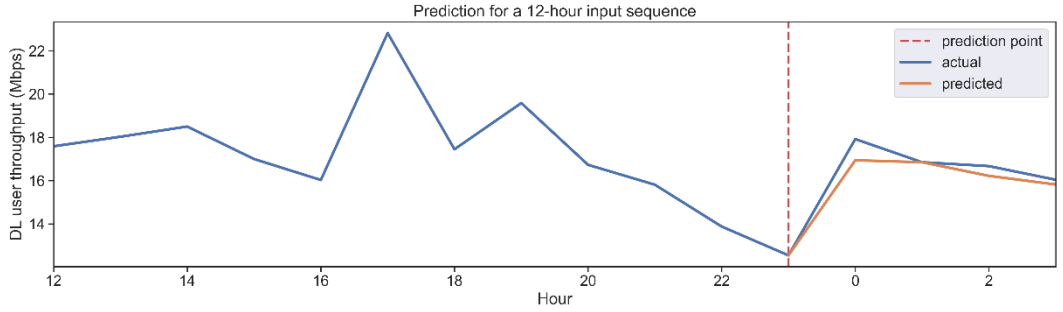

Figure 4.11 Temporal throughput prediction for one cell (12-hour observation)

To further understand the goodness of the model across the entire region, the MAPE for the 100 grids at the third hour of prediction is calculated. The obtained result is illustrated in. The interpretation of the results is based on the definition mentioned in Table 4.4.

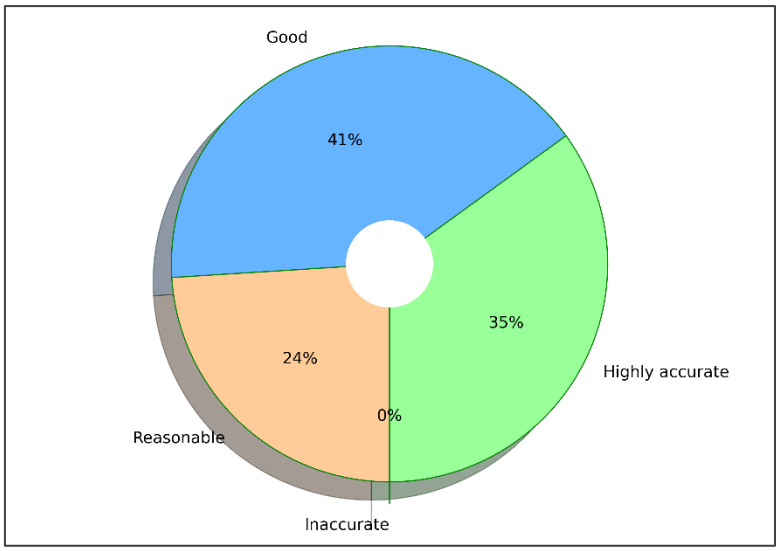

Figure 4.12 MAPE distribution over 100 grids.

The spatial prediction result for the whole target area at the third hour for 24-hour and 12-hour input sequence models are shown in Figure 4.13 and Figure 4.14, respectively. In both figures, the image on the left represents the actual throughput values, while the right one is the prediction made by the model. From the two figures, we can see that the prediction result with a 12-hour input captures the patterns of the original data better than the 24-hour input. Hence, the model with a 12-hour input shape performs better in spatial dimensions as well.

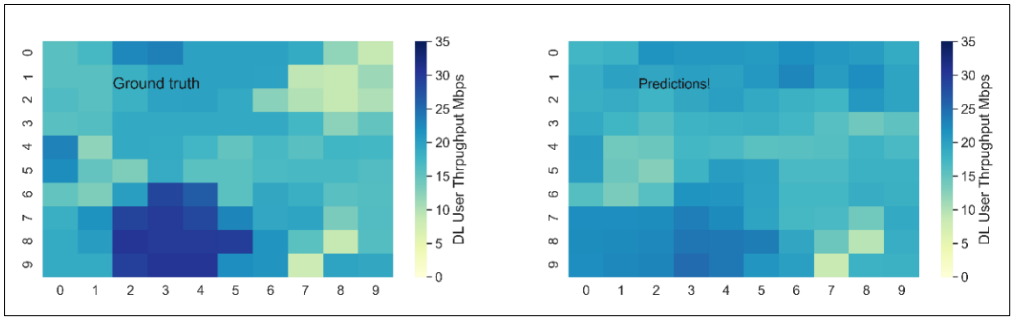

Figure 4.13 Spatial throughput prediction after 3 hours (24-hour observation)

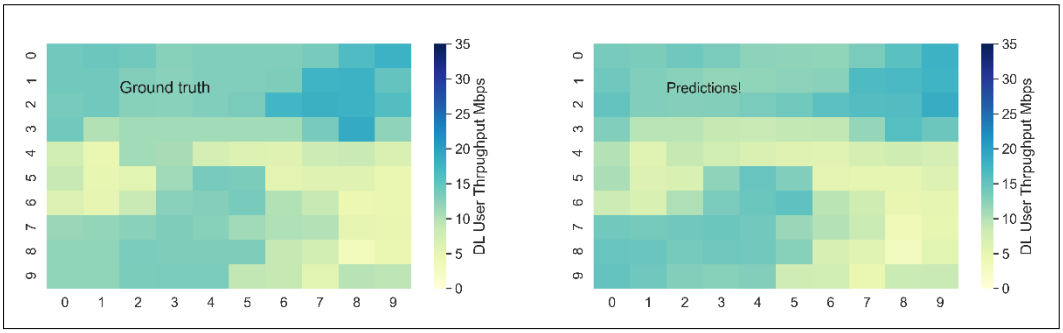

Figure 4.14 Spatial throughput prediction after 3 hours (12-hour observation)

4.8 Detecting Poor Coverage

To identify areas with poor coverage, the coverage analysis approach mentioned in Section 4.3 is employed. First, the average signal strength and coverage probabilities of cells serving each geo-grid have been calculated. Next dominant cells in each grid are identified (serving cells with coverage probability > 0.5 are said to be 'dominant cells'). The grid-level coverage map of the target area along with the dominant serving cells is depicted in Figure 4.15. Annotations in each grid depict the dominant cell.

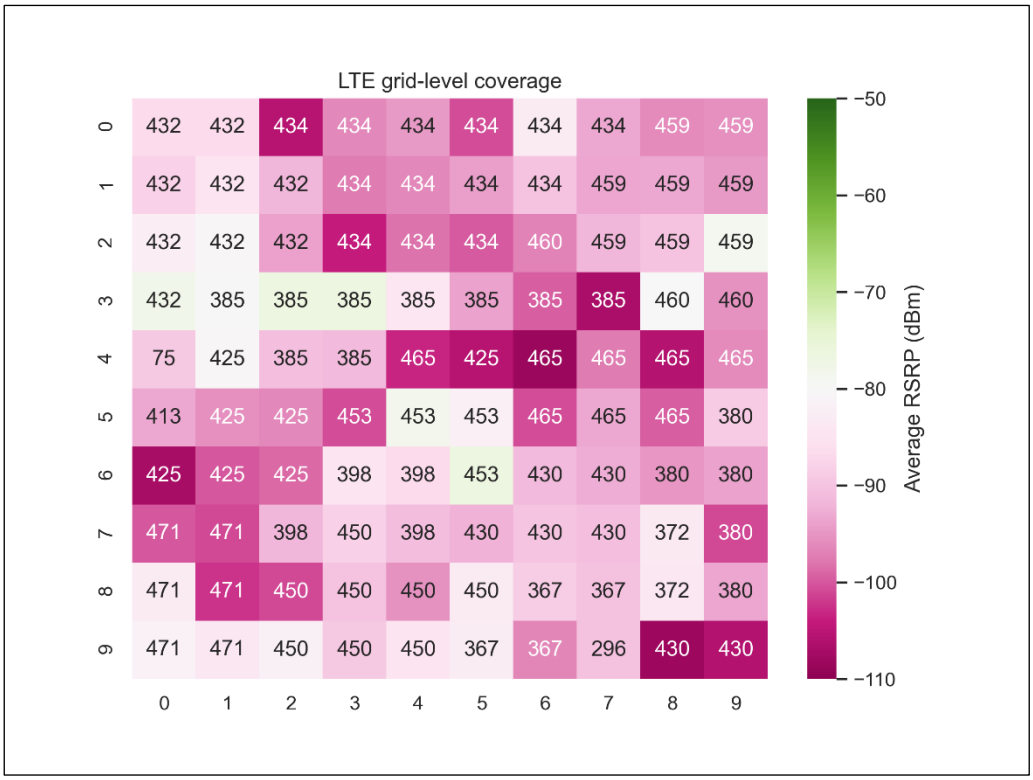

Figure 4.15 LTE coverage map with dominant cells

It's evident from Figure 4.15 that areas with gird numbers '06', '20', '32, '64','73','89','94', and 99 have the worst coverage with a mean RSRP close to -110 dBm.

4.9 Throughput Distribution Map

To obtain LTE average DL user throughput map, the grid level throughput data obtained in Section 4.5.2 is utilized. Network busy hour performance data is commonly used in network planning and optimization activities. As a result, network busy hour grid throughput data is used to investigate average user throughput performance in various regions of the target area. Figure 4.16 depicts the throughput distribution map.

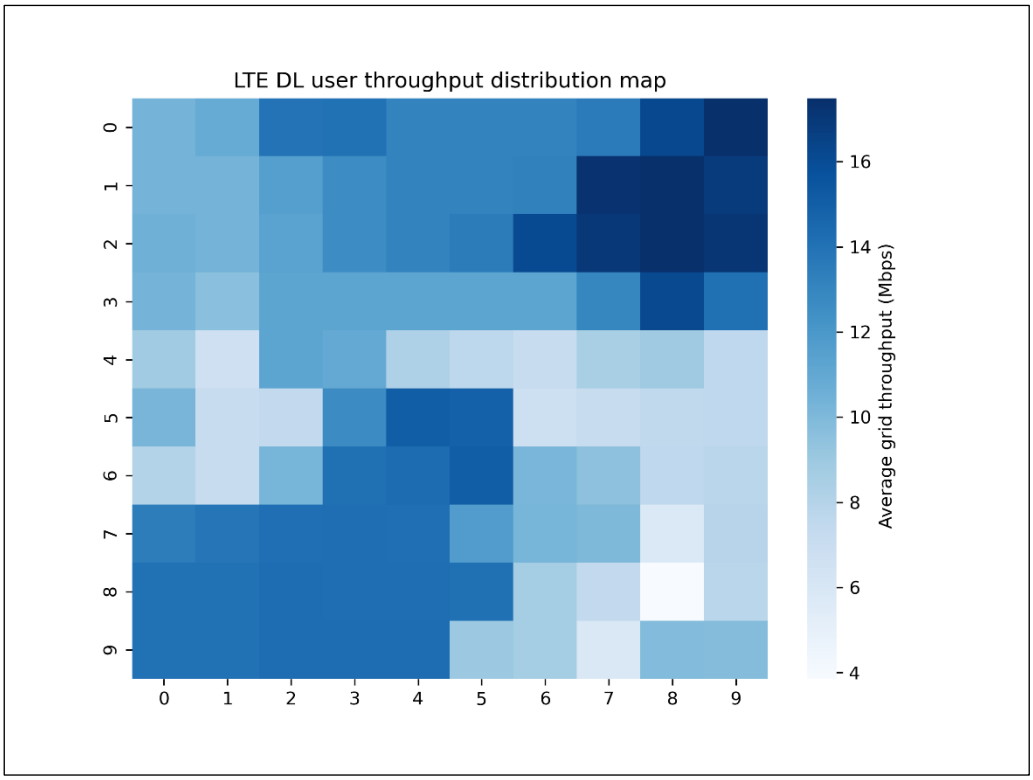

Figure 4.16 LTE average DL user throughput distribution map

Figure 4.16 clearly shows that the bottom right corner and mid-left regions of the target area show relatively lower average user throughput with a value below 10 Mbps.

5 Conclusion and Future Directions

5.1 Conclusion

In this work, spatiotemporal modeling and prediction of LTE DL user throughput are conducted using the ConvLSTM DNN algorithm. For spatiotemporal modeling, a technique that fuses realistic coverage measurement and average DL user throughput data was implemented. From this technique, spatiotemporal average DL user throughput data with a spatial grid resolution of 100m by 100m was obtained. Furthermore, the modeling results have been used to locate poor coverage areas, and analyze throughput distribution across the selected region.

The developed model was able to predict average user throughput 3 hours in the future with RMSE of 2.02, MAE of 1.52, and MAPE of 14.68. Moreover, MAPE distribution across the different grids is analyzed to get a good picture of how well the model generalizes across the whole region. 41%, 35%, and 24% of the grids showed highly accurate, good, and reasonable predictions, respectively. On the other hand, no grids showed inaccurate predictions.

The findings from this study can be used for proactive QoE assurance, specifically throughput demands. On top of that, the approaches can be applied for better planning and optimization of mobile data networks. This conclusion is reached since the study has been able to achieve:

- Throughput distribution with as little as 100m grid resolution;
- Coverage map with a resolution of the above grid size;
- Serving cells dominating each grid;
- Accurate prediction of user throughput for 3 hours.

Furthermore, knowledge of throughput distribution and coverage map in a fine grid resolution together with the dominating cells could be an input for further root cause analysis.

5.2 Future Directions

Some interesting future research directions can be drawn from this study. First, since this work is based on the assumption of uniform user distribution, it would be interesting to study realistic user distribution. This can be achieved by studying the land-use class within the target area (e.g., industrial, commercial, small building, etc.) and estimating the distribution of users based on the day of the week and the time of the day. Realistic user distribution would allow us to accurately estimate throughput distribution across spatial grids.

Secondly, crowdsourcing can be used to enhance the drive test approach implemented in this work. This could be quite helpful, especially in indoor areas, which are inherently unreachable by drive test. By fusing data from the two sources, a higher resolution grid can be formed. Spatial interpolation techniques might also be used to fill areas where measurement is not available.

It would also be interesting to include features that affect throughput, such as channel quality indicator and active number of users. Finally, other architectures of ConvLSTM such as encoder-decoder can be explored.

References

[1] J. Riihijarvi and P. Mahonen, "Machine Learning for Performance Prediction in Mobile Cellular Networks," *IEEE Computational Intelligence Magazine*, vol. 13, no. 1, pp. 51–60, Feb. 2018, doi: 10.1109/MCI.2017.2773824.

[2] *Universal Mobile Telecommunications System (UMTS); LTE; Telecommunication management; Key Performance Indicators (KPI) for Evolved Universal Terrestrial Radio Access Network (E-UTRAN): Definitions*, 3rd Generation Partnership Project (3GPP), Technical specification (TS) 3GPP, document 32.450, version 9.3.0, Oct. 2011.

[3] T. ur Rehman, M. A. I. Baig, and A. Ahmad, "LTE downlink throughput modeling using neural networks," in *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, Oct. 2017, pp. 265–270. doi: 10.1109/UEMCON.2017.8249044.

[4] "xMbps Anytime Anywhere White Paper", Huawei Technologies Co., Ltd., 2019, Accessed: Sep. 30, 2021. [online]. Available: https://www-file.huawei.com/-/media/corporate/pdf/publications/communicate/75/07-focus--anytime-anywhere%20xmbps%20building%20quality%20wireless%20based%20on%20user%20perception.pdf.

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

[6] C. Zhang and P. Patras, "Long-Term Mobile Traffic Forecasting Using Deep Spatio-Temporal Neural Networks," in *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, New York, NY, USA, Jun. 2018, pp. 231–240. doi: 10.1145/3209582.3209606.

[7] Geleta Abdisa, "UMTS Network Coverage Hole Detection using Decision Tree Classifier Machine Learning Approach," Master Thesis, Addis Ababa Institute of Technology, Addis Ababa, Feb. 2020.

[8] X. Zuo, "A Deep Learning Approach to Downlink User Throughput Prediction in Cellular Networks," Master Thesis, KTH Royal Institute of Technology, Stockholm, 2020.

[9] Demisse Hailemariam, "Spatiotemporal Mobile Data Traffic Prediction Using Convolutional Long Short-Term Memory: The case of Addis Ababa, Ethiopia," Master Thesis, Addis Ababa Institute of Technology, Addis Ababa, Dec. 2019.

[10] 3GPP, "LTE Specification," [Online]. Available: http://www.3gpp.org/technologies /keywords-acronyms/98-lte. [Accessed: 30-Sep-2021]. Apr. 2010.

[11] A. Elnashar, M. A. El-saidny, and M. R. Sherif, *Design, deployment and performance of 4g-lte networks: A practical approach*. Chichester, England: Wiley, 2014.

[12] H. Holma, A. Toskala, and P. Tapia, *HSPA+ evolution to release 12: Performance and optimization*. Chichester, England: Wiley, 2014.

[13] Y. CHEN and X. Lagrange, "Architecture and Protocols of EPC-LTE with relay," 2013. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00830621

[14] Atayero, A. A. and Luka, M. K. and Orya, M. K and Iruemi, J. O., "3GPP Long Term Evolution: Architecture, Protocols and Interfaces," *International Journal of Information and Communication Technology Research*, vol. 1, no. 7, pp. 306-310, Nov. 2011.

[15] V. Buenestado, J. M. Ruiz-Avilés, M. Toril, S. Luna-Ramírez, and A. Mendo, "Analysis of Throughput Performance Statistics for Benchmarking LTE Networks," *IEEE Communications Letters*, vol. 18, no. 9, pp. 1607–1610, Sep. 2014, doi: 10.1109/LCOMM.2014.2337876.

[16] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin, "Understanding Mobile Traffic Patterns of Large Scale Cellular Towers in Urban Environment," *IEEE/ACM Transactions on Networking*, vol. 25, no. 2, pp. 1147–1161, Apr. 2017, doi: 10.1109/TNET.2016.2623950.

[17] J. F. Torres, D. Hadjout, A. Sebaa, F. Martínez-Álvarez, and A. Troncoso, "Deep learning for time series forecasting: A survey," *Big Data*, vol. 9, no. 1, pp. 3–21, 2021.

[18] P. Hendikawati, Subanar, Abdurakhman, and Tarno, "A survey of time series forecasting from stochastic method to soft computing," *J. Phys.: Conf. Ser.*, vol. 1613, p. 012019, Aug. 2020, doi: 10.1088/1742-6596/1613/1/012019.

[19] M. Joshi and T. H. Hadi, "A Review of Network Traffic Analysis and Prediction Techniques," *arXiv:1507.05722 [cs]*, Jul. 2015, Accessed: Sep. 29, 2021. [Online]. Available: http://arxiv.org/abs/1507.05722

[20] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, Jan. 2003, doi: 10.1016/S0925-2312(01)00702-0.

[21] Y. Abera and D. Hailemariam, "Spatio-temporal Mobile Data Traffic Modeling Using Fourier Transform Techniques," in *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, Oct. 2018, pp. 20–24. doi: 10.1109/ICTC.2018.8539698.

[22] L. Guan, X. Zhang, Z. Liu, Y. Huang, R. Lan, and W. Wang, "Spatial modeling and analysis of traffic distribution based on real data from current mobile cellular networks," in *2013 International Conference on Computational Problem-Solving (ICCP)*, Oct. 2013, pp. 135–138. doi: 10.1109/ICCPS.2013.6893524.

[23] J. Hurwitz, and D. Kirsch, *Machine Learning For Dummies*. Chichester, England: Wiley, 2018

[24] J. Wang *et al.*, "Spatiotemporal modeling and prediction in cellular networks: A big data enabled deep learning approach," in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, 2017, pp. 1–9.

[25] J. Brownlee, *Long short-term memory networks with Python: develop sequence prediction models with deep learning*. Australia: Machine Learning Mastery Pty. Ltd, 2018.

[26] I. Vasilev, D. Slater, G. Spacagna, P. Roelants, and V. Zocca, *Python deep Learning: Exploring deep learning techniques and neural network architectures With Pytorch, Keras, and TensorFlow*. Birmingham: Packt Publishing, 2019.

[27] J. Brownlee, "A Gentle Introduction to Pooling Layers for Convolutional Neural Networks," *Machine Learning Mastery*, 05-Jul-2019. [Online]. Available: https://machinelearningmastery.com/pooling-layers-for-convolutional-neural-networks. [Accessed: 30-Sep-2021].

[28] J. Sun and J. Kim, "Joint prediction of next location and travel time from urban vehicle trajectories using long short-term memory neural networks," *Transportation Research Part C: Emerging Technologies*, vol. 128, p. 103114, Jul. 2021, doi: 10.1016/j.trc.2021.103114.

[29] Dziugas Vysniauskas, "Spatio-Temporal Mobile Traffic Forecasting," Master Thesis, University of Edinburgh, Edinburgh, 2020.

[30] Zeneb Kassaw, "Coverage Prediction Based on Spatial Interpolation Techniques: The Case of UMTS Network in Addis Ababa, Ethiopia," Master Thesis, Addis Ababa Institute of Technology, Addis Ababa, Jan. 2020.

[31] H. Wang, S. Xie, K. Li, and M. O. Ahmad, "Big Data-Driven Cellular Information Detection and Coverage Identification," *Sensors*, vol. 19, no. 4, Art. no. 4, Jan. 2019, doi: 10.3390/s19040937.

[32] B. B. Haile, E. Mutafungwa, and J. Hämäläinen, "A Data-Driven Multiobjective Optimization Framework for Hyperdense 5G Network Planning," *IEEE Access*, vol. 8, pp. 169423–169443, 2020, doi: 10.1109/ACCESS.2020.3023452.

[33] D. M. Rose, J. Baumgarten, and T. Kurner, "Spatial Traffic Distributions for Cellular Networks with Time Varying Usage Intensities Per Land-Use Class," in *2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, Sep. 2014, pp. 1–5. doi: 10.1109/VTCFall.2014.6966182.

[34] G. Barlacchi, M. De Nadai, R. Larcher, A. Casella, C. Chitic, G. Torrisi, F. Antonelli, A. Vespignani, A. Pentland, and B. Lepri, "A multi-source dataset of urban life in the city of Milan and the province of Trentino," *Scientific Data*, vol. 2, no. 1, 2015.

[35] S. Bhanja and A. Das, "Impact of Data Normalization on Deep Neural Network for Time Series Forecasting," *arXiv:1812.05519 [cs, stat]*, Jan. 2019, Accessed: Sep. 29, 2021. [Online]. Available: http://arxiv.org/abs/1812.05519.

[36] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *International Conference on Machine Learning*, Jun. 2015, pp. 448–456.

[37] K. S. V. Muralidhar, "Learning curve to identify overfitting and Underfitting in machine learning," *Medium*, 22-Feb-2021. [Online]. Available: https://towardsdatascience.com/learning-curve-to-identify-overfitting-underfitting-problems-133177f38df5. [Accessed: 30-Sep-2021].

[38] A. Jierula, S. Wang, T.-M. Oh, and P. Wang, "Study on Accuracy Metrics for Evaluating the Predictions of Damage Locations in Deep Piles Using Artificial Neural Networks with Acoustic Emission Data," *Applied Sciences*, vol. 11, no. 5, Art. no. 5, Jan. 2021, doi: 10.3390/app11052314.