ADDIS ABABA UNIVERSITY

ADDIS ABABA INSTITUTE OF TECHNOLOGY

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

Spatiotemporal Mobile Data Traffic Prediction Using Convolutional Long Short-Term Memory: The case of Addis Ababa, Ethiopia

By

Demisse Hailemariam

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 Masters of Science in Telecom Information System

December, 2019

Addis Ababa, Ethiopia

ADDIS ABABA INSTITUTE OF TECHNOLOGY

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

TELECOMMUNICATION ENGINEERING GRADUATE PROGRAM

Spatiotemporal Mobile Data Traffic Prediction Using Convolutional Long Short-Term Memory: The case of Addis Ababa, Ethiopia

By

Demisse Hailemariam

Dr. –Ing. Dereje Hailemariam	_____
Advisor	Signature
_____	_____
Internal Examiner	Signature
_____	_____
External Examiner	Signature

Declaration

I, the undersigned, declare that this thesis is my original work, has not been presented for a degree in this or any other university, and all sources of materials used for the thesis have been fully acknowledged.

Demisse Hailemariam _____

Name Signature

Place: Addis Ababa Institute of Technology, Ethiopia

Date of Submission: _____

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

Dr. –Ing. Dereje Hailemariam _____

Advisor Signature

Abstract

Globally, exponential data growth is observed with mobile traffic generated from devices like tablets, smartphones and other devices. Likewise, Addis Ababa city's cellular network data traffic is increasing exponentially. To absorb this high traffic demand ethio telecom, the telecom service provider in the city continuously expands and optimizes the cellular network. Having knowledge of the growing data traffic demand in advance at a given time and space will assist ethio telecom's planning strategy and optimization.

Recently, few studies are conducted to forecast Addis Ababa city's Universal Mobile Telecommunication System (UMTS) network traffic using statistical time series models and neural network models. However, the studies deal with only time-domain forecasting and recommend to do from a spatial point of view. Moreover, another study modeled the spatiotemporal mobile data traffic, which can capture the space and time variation of UMTS data traffic in the city; the study recommends the need of spatiotemporal data traffic prediction.

In this thesis, a deep neural network model, specifically Convolutional Long Short-Term Memory (ConvLSTM), is used for spatiotemporal data traffic demand prediction of Addis Ababa city. Three months' real dataset from 739 base stations is collected and preprocessed from ethio telecom's UMTS network. After defining geographical grids, the ConvLSTM model is applied, which can capture spatial correlations through convolution operators and temporal dynamics through the LSTM network for prediction.

The proposed model can predict up to six hours of future data traffic with a root mean square error (RMSE) of 1.37. Additionally, the predicted data traffic demand is analyzed with respect to blocked data traffic at a given space and time which gives significant insight to the optimization processes like load balancing.

Keywords – Spatiotemporal, data traffic prediction, UMTS network, deep neural network, ConvLSTM.

Acknowledgment

This would not have been possible without the help of GOD. I just want to say praise to Jesus Crist and his Mother Holy Virgin Mariam for the endless help throughout my life.

My first and foremost thanks goes to Dr. -Ing. Dereje Hailemariam for his tremendous support, patience, encouragement and enthusiasm. He has been very dedicated throughout my thesis. I would also like to thank Ms. Bethelhem Seifu for her unreserved effort for the betterment of this thesis.

I am also so grateful to thank my colleagues, especially Ato Habtmu Abayneh, Ato Buzayehu Zerhune and Ato Mesfin Abatkun for their great collaborations and encouragements throughout the whole process.

Finally, I would like to thank my deepest appreciation to my better half Berhane Alemayehu, for all the love and positive mentality that she has given me throughout the time of my research.

Table of Contents

List of Figures

List of Table

Abbreviations

2D	Two-dimension
3D	Three-dimension
3D CNN	Three-dimensional Convolutional Neural Network
3G	Third Generation
4G	Fourth Generation
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
CAGR	Compound Annual Growth Rate
CAPEX	Capital Expenditure
CDR	Call Detail Record
CNN	Convolutional Neural Network
ConvLSTM	Convolutional Long Short-Term Memory
GB	GigaByte
GSM	Global System for Mobile communication GSM
HSDPA	High-speed downlink packet access
LTE	Long Term Evolution
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MLP	Multilayer Perceptron

OPEX	Operational Expenditure
PS	Packet switched
RAB	Radio Access Bearer
RF	Radio Frequency
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
SARIMA	Seasonal Autoregressive Integrated Moving Average
SARIMA-ELM	SARIMA-Extreme Learning Machine
TB	TeraByte
UMTS	Universal Mobile Telecommunication System

1. Introduction

The ever-changing mix and growth of wireless devices that are accessing mobile networks worldwide are some of the primary contributors to global mobile traffic growth. Global mobile data traffic is expected increase seven-fold between 2016 and 2021 with a compound annual growth rate (CAGR) of 47 percent from 2016 to 2021, reaching 49.0 Exabytes per month by 2021 [1].

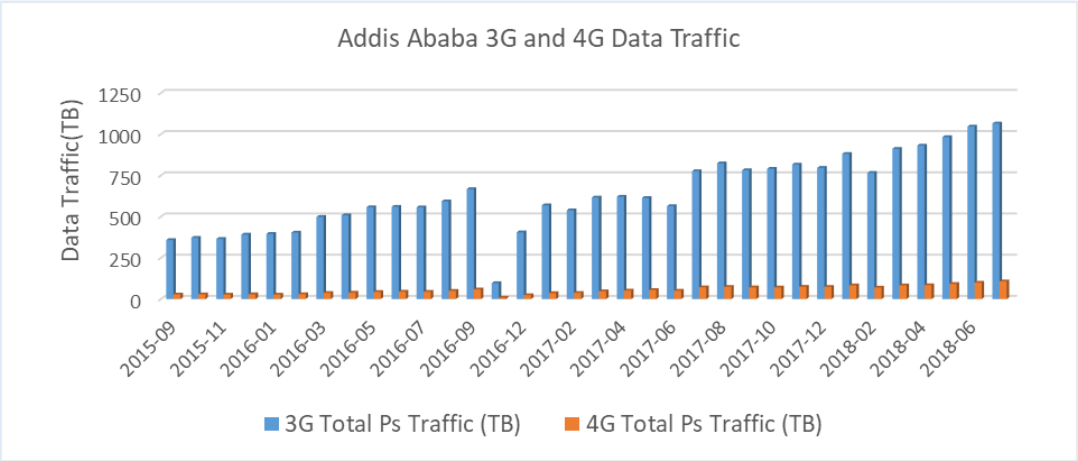


Figure 1-1 Monthly data traffic growth trend for Addis Ababa [2].

ethio telecom's Addis Ababa cellular network data traffic growth also follows the same increasing pattern. As shown in Figure 1-1, the monthly traffic for the city is steadily increasing for both 3G and 4G technologies with the data volume download reaching 1,140TB per month in 2018. For the past three years, CAGR of the city's data traffic is on the average 39 percent, which is close to the global data growth trend. For this growing data traffic, knowing the forecasted data traffic demand is crucial.

Currently, ethio telecom deployed 332 Long Term Evolution (LTE) and 739 UMTS sites to support the data traffic demand of Addis Ababa. To improve the user's experience and absorb the data traffic growth, ethio telecom implemented different mechanisms, such as load balancing by parameter optimization, resource expansion like carrier addition or board expansion, and base station densification on a greenfield or rooftop. ethio telecom uses current data traffic demand,

customer complaints, and marketing information as an input for planning, optimization, and management of the network.

For network operators such as ethio telecom, proactively knowing the demand growth will facilitate the network expansion and resource allocation process before the customer experience the network congestion, and helps to deliver a better quality of services. Prediction of mobile data traffic is also important for managing resources in advance, minimizing operational and maintenance costs, and satisfy user's needs [3].

Data traffic prediction can be characterized in time, space or both aspects. In the temporal domain, its periodicity and seasonality of data traffic are studied whereas in the spatial domain, the data traffic is discussed based on geography or space. But in the spatiotemporal domain, both characteristics which are found in time and space domain are discovered [4]

Prior studies are done on data traffic forecasting for Addis Ababa UMTS network from the time domain viewpoint. In [5], both Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average (SARIMA) models were investigated as alternative ways of forecasting data traffic by using real mobile data collected from the operator's network. Both SARIMA and ARIMA are linear models that will not capture the non-linearity prevailing in data traffic. The performance of hybrid SARIMA and Extreme Learning Machine (ELM) based data traffic forecasting for Addis Ababa UMTS network is investigated studied in [6]. ELM is one variant of Artificial Neural Network and predicts the non-linearity part of the data traffic, which is used to improve the prediction by the SARIMA, which in turn captures the linear part of the data. However, the prior studies are dealing only from a temporal perspective.

In addition to this, spatiotemporal mobile data traffic modeling is presented by [7], which can capture the spatiotemporal dynamics of Addis Ababa UMTS mobile data traffic, and recommend spatiotemporal prediction as to future work. The main motivation of this research steams from these prior researches and to apply deep neural network techniques to capture the time and space dependence of the data traffic and further predict the traffic demand. Applying deep neural network models leads to better prediction accuracy [8]. In this thesis, Addis Ababa city

mobile data traffic demand is discovered from a temporal and spatial point of view and a spatiotemporal data traffic prediction is implemented using deep neural networks.

1.1. Statement of the Problem

Addis Ababa city ethio telecom cellular network data traffic demand is growing day by day. This brings the company opportunities, from a revenue perspective, and challenges, as it puts a burden in the network and users expect data service with acceptable quality. Supporting the data requires availing infrastructure and network management solutions and also efficient utilization of key resources like spectrum and energy.

Currently, to handle this growing data traffic demand, planning and management of the network are done, and it relies on the utilization of the network resource and current traffic demand. To address the requirements of rapidly growing data traffic demand with a suitable plan, applying the current data traffic as an input is not sufficient, instead accurately predicted data traffic demand at a given time and space is needed. However, managing data traffic is complex, due to its exceptional temporal and spatial dependencies [4].

Understanding the dynamic behavior of mobile data traffic and forecast future demand of it provides extra power to manage and control the outburst data traffic by network planning, data pricing, and load balancing mechanisms [9]. Using state-of-the-art approaches, such as one variant of deep neural network termed Convolutional Long Short-Term Memory (ConvLSTM) is expected to improve the spatiotemporal prediction accuracy.

1.2. Objective

1.2.1 General Objective

The main objective of this thesis is to apply ConvLSTM deep neural networks to model and predict the spatiotemporal variation of the UMTS mobile data traffic in the city of Addis Ababa.

1.2.2 Specific Objectives

To achieve the general objective of the study the following specific objectives are identified.

- Explore related works on spatiotemporal data traffic prediction using literature reviews.
- Collect and pre-process the data.
- Prepare and analyze dataset for prediction by selecting a target area and define a regular grid (from space and time perspective).
- Predict spatiotemporal data traffic of Addis Ababa UMTS network using deep neural network (ConvLSTM).
- Evaluate the performance of the model using the performance evaluation metrics, such as RMSE, and select the best fit model based on minimum prediction error by tuning hyperparameters of the model.

1.3. Literature Review

Related works have been studied on spatiotemporal data traffic prediction using different approaches. In [10] spatiotemporal individual mobile data traffic prediction is investigated. The focus is predicting the mobile data traffic generated among each user who consumes a large amount of data traffic. To study the temporal and spatiotemporal dynamics two prediction methods namely, Markovian predictors method and Multilayer Perceptron (MLP) have been used. The investigation is done on the basis of forecasting scenarios of data traffic volume in isolation and study of predictability of mobile data traffic per-users in visited locations.

According to [11], the importance of forecasting in high accuracy the volume of data traffic that mobile users will consume is highly increasing. The authors objectively designed network-wide mobile traffic forecasting to solve the challenges through deep learning and propose Double Spatio-temporal neural network (D-STN) architecture. The proposed solution performs up to 10-Hours long prediction. The authors used ConvLSTM and three-dimensional Convolutional Network (3D-ConvNet) structures to model long term trends and short-term variations of the mobile traffic volume, and finally, MLP is employed to map the output on the final prediction.

Multitask learning architecture using deep neural networks for mobile traffic forecasting is done to take full advantage of machine learning [8]. In data recording, a multi-source data set from Millan is used. While the authors divided the city into 100 * 100 areas, geographical grids are first defined with aggregated call detail record (CDR) data called Milan Grid. By observing traffic data in the last hour, the models are designed to predict the maximum, average, and minimum traffic loads in the next hour. The combination of Recurrent Neural Network and Convolutional Neural Network (RNN-CNN) is carefully used to take advantage of related outputs in multitask learning. The experiment reveals that CNN and RNN can extract geographical and temporal traffic features respectively.

The study in [12] discovers that due to the tremendous temporal and spatial dynamics which introduce by diverse user internet behavior and frequent user mobility all over the city, the prediction of cellular traffic is highly challenging in the modern metropolis. The authors who are motivated by the decomposition of in cell and inter-cell data traffic, and using a directed graph to model the Spatio-temporal features of traffic patterns. They proposed a graphical neural network (GNN) that can learn from a graphic structure. Based on big cellular usage data set covering 1.5 million users, large scale and fine-grained predictions were performed for half an hour traffic. The experiments show that the significance of the interaction of spatial and temporal factors for accurate prediction.

To summarize, for the implementation of spatiotemporal mobile data traffic demand prediction, deep learning models are more preferred than time series prediction models [11], similar to the one followed in this thesis. This is mainly because of their capability to capture the spatial and temporal dynamics of the dataset. Especially, CNN helps to get the spatial feature of the dataset and applying LSTM helps to get time-series dynamics of the dataset. Besides, several works focus on the usage traffic collected from CDR or measurement reports, without considering the blocked traffic, which is the traffic demand not served by the network due to insufficient resources.

1.4. Methodology

Related works on spatiotemporal data traffic prediction investigated and the historical dataset is collected from ethio telecom's UMTS network accordingly. In the first phase, the dataset is prepared for analysis to capture space and time dynamics of the network after selecting the target area and define a regular grid. Then, the dataset is split into training and test set. Finally, the model is trained and tested with a performance evaluation metrics until better prediction accuracy is achieved. In the training process, the hyperparameters are tuned to minimize the prediction error. To simulate and evaluate the possible results of the study the following tools such as Microsoft Excel, MapInfo, Python, Keras and TensorFlow libraries are used.

1.5. Scope and Limitation

1.5.1 Scope

This research is focused on Addis Ababa city's UMTS network spatiotemporal data traffic prediction based on base station level dataset, even if the UMTS network has voice and data services.

1.5.2 Limitation

The research has some limitations such as

- The spatial data traffic distribution below base station level is not covered in this study, because of the limitation of finding historical mobile data traffic measurement for such kind of spatial data traffic distribution in small grid resolution.
- Due to the complexity of the computational process of the model, the study is conducted on the selected area from Addis Ababa city's UMTS network.

1.6. Contribution

This research output provides significant insight to ethio telecom Addis Ababa's growing data traffic demand and helps to plan the effective cellular network deployment strategy with the consideration of this forecasted data traffic as an input. Moreover, proactively knowing the data traffic demand at a given time and space helps to select hotspot areas, and also, it can be used as an input for proper optimization, network management, and dynamic data pricing process.

1.7. Thesis Outline

The rest of the paper is structured as follows. Chapter 2 describes the basics of UMTS technology and data traffic demand, Chapter 3 introduces the mobile data traffic characteristics and prediction. The chapter also describes the need for a deep neural network for data traffic prediction. Chapter 4 presents the result and discussion part, first experimental setups for the data traffic prediction is presented, and then the result of the prediction models and findings are discussed. Finally, Chapter 5 presents the conclusion and future works.

2. Universal Mobile Telecommunication System

In this Chapter, the basic concepts of UMTS network and its network architecture are explained. Moreover, the Chapter discusses services that are delivered by UMTS network, and specifically on mobile data traffic demand which is the total traffic offered by the network.

2.1. Overview of UMTS Network

Mobile network technologies advancement is increasing from generation to generation to deliver a better quality of voice and data service, with affordable cost for the user and minimum capital expenditures (CAPEX) and operating expenses (OPEX) from the operator side. UMTS is the third-generation technology (3G) that is implemented based on the Global System for Mobile communication (GSM). In the second-generation (2G) evolution process, GSM/General Packet Radio Service (GPRS) is evolved after GSM implemented. GSM/GPRS technology can deliver services using a packet switched in addition to the circuit switched. From GPRS the extended technologies Enhanced GPRS (EGPRS)/Enhanced Data rates for GSM Evolution (EDGE) is advanced, then after the UMTS (3rd generation network) is evolved [13].

UMTS provides data and voice services, and support features like internet and video streaming with a speed of 2Mbps data transfer rate. UMTS network is deployed on top of wideband code data multiple access (WCDMA) technology and it is upgraded to High-speed packet access (HSPA). In the case of HSPA, the expected pick data rate from High-speed downlink packet access (HSDPA) and High-speed uplink packet access (HSUPA) is up to 7.2Mbps and 4Mbps respectively [14]. When deploying UMTS on top of 2100MHz the coverage setup is 1-1.5km, so it needs more base station than GSM network for large coverage area. But the coverage issue can be handled by applying 850 MHz or 900 MHz band that have the potential to cover long-distance [13].

2.2. UMTS Network Architecture

UMTS network architecture is most like GSM/GPRS network architecture because it is based on the GSM standards. The network elements found in the UMTS system are grouped into three

subsystems according to their functionalities. The first one is the User Equipment (UE), that works as an interface between users and radio access network.

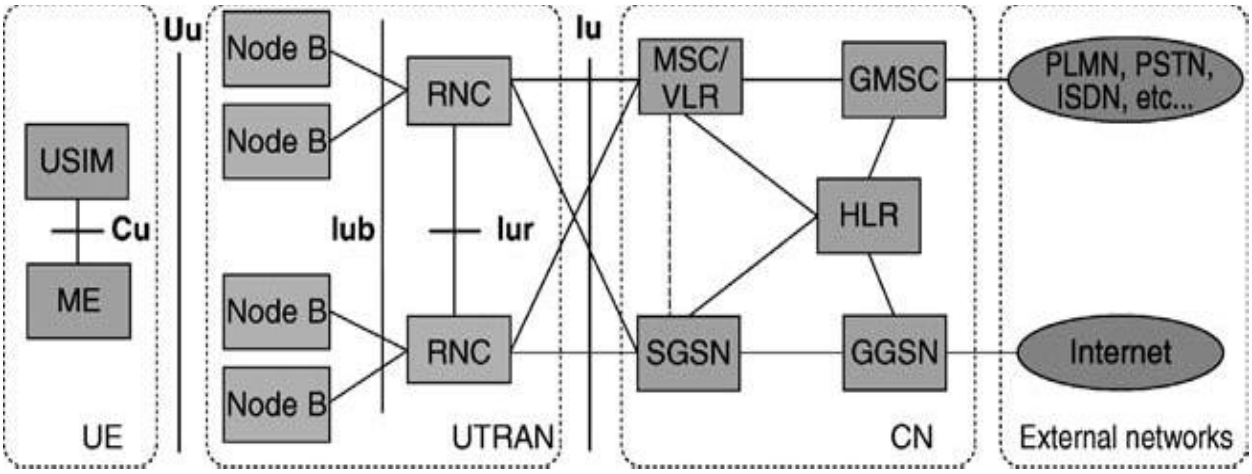


Figure 2-1 UMTS Network Architecture [15].

UMTS Terrestrial Radio Access Network (URTAN) is also responsible for controlling radio-related functionalities and the last one is Core Network (CN), which handles the interaction with the external network in addition to switching and routing calls, as shown in Figure 2-1. The main purposes of the Network elements are discussed below [15].

2.2.1 User Equipment

The UE consists of two components, Mobile Equipment (ME) which is a terminal used for radio communication and UMTS Subscriber Identity Module (USIM) smartcard with subscriber identity which used to store some subscriber information and handle the tasks related to it [15].

2.2.2 UMTS Terrestrial Radio Access Network

UTRAN consists of two network elements, Node B and Radio Network Controller (RNC). Nod B is similar to the base station in the GSM network, it provides voice and data coverage for UE, and control the air interface (the radio link between UE and Nod B). Whereas RNC is used to manage the Node B elements and controls radio resources, and it provides services requested by CN [15].

2.2.3 Core Network

In Core Network there are two major domains to be discussed. These are circuit switched domain and packet switched domain. Mainly, Circuit switched is dedicated resources to each user like telephone service, and packet switched shares resources with multiple users and used to deliver packet services such as the internet. The components of each domain are described follows [15]

- Circuit switched domain components
  - (Mobile Services Switching Centre/Visitor Location Register (MSC/VLR): - The MSC is used to control the circuit switched transactions like voice call and text messaging setup, and VLR serves as a database for UE that currently visits the specific location.
  - Gateway MSC (GMSC): - is used to provide a connection to the external network.
- Packet switched domain components
  - Serving GPRS Support Node (SGSN): - It has similar functionalities as MSC/VLR but it used for setup and management of data connection between the packet switched and UE
  - Gateway GPRS Support Node (GGSN): - has similar functionalities as GMSC but it is working on packet switched service.

Home Location Register (HLR) is one of the main components of CN which is a database that stores all information related to users. When the user joins the network for the first time, it will register to HLR and it will stay as far as it is an active user. And when the UE connects to the MSC/VLR in the circuit switched domain or SGSN in the packet switched domain, the user information is copied to them from HLR [15].

2.2.4 UMTS Network Interfaces

UMTS network facilitates communication between the above subsystems using different interfaces, some of them are described as follows [15].

I. Cu interface: - connects the ME to USIM smart card.
II. Uu interface: - the radio link between Node B and the UE, it is called as “air” interface.

III. Iu interface: - connects the CN to the UTRN, specifically IuCS interface connects the circuit switched network to the UTRAN and IuPS interface connects the packet switched network to the UTRAN

IV. Iub interface: - connects RNC to multiple Node B.

V. Iur interface: - provides the connection between RNC and allows soft handover between them.

2.3. UMTS Services and Data Traffic Demand

Telecom operators deploy UMTS network to provide different types of voice and data services. In doing so, the UTRAN subsystem assigned dedicated resources for transportation of these services, which is named as radio access bearer (RAB). In the following subsections, RAB services are described and data traffic demand is elaborated accordingly.

2.3.1 Radio Access Bearer

RAB is a service provided by UTRAN for UE and CN. It is described by a set of parameters for a specific traffic class. Various RAB types are defined by the 3rd Generation Partnership Project (3GPP) considering transport channel parameters for each RAB type and different RAB combinations on uplink and downlink for transferring different traffic classes. The detailed transport channel parameters of the given RAB type are defined by service type which is also referred as traffic class, Peak rate (in Kilobits per second (Kbps)), Payload (in bit), Number of Transport blocks, Transmission Time Interval (milliseconds) and Radio Link Control (bit). For a particular connection, RNC selects the appropriate RAB according to the parameters provided by SGSN during RAB establishment and transfers data with a certain quality of service [16].

Table 2-1 Traffic Class [16]

Traffic class	Fundamental characteristics
Conversational class	▪ Require a certain reservation of resources in the network. ▪ It is delay sensitive, which is intended for carrying real-time traffic flow. ▪ Example: speech, video
Streaming class	▪ Require a certain reservation of resources in the network. ▪ Preserve time variation between information entities of the stream (i.e. some but constant delay) ▪ Example: streaming audio and video
Interactive class	▪ No need of resource reservation and throughput depends on the load in the cell. ▪ It is request response pattern ▪ Example: web browsing
Background	▪ No need of resource reservation and throughput depends on the load in the cell. ▪ The expectation on the destination side is not at a certain time. ▪ Example: background download of emails or file downloads

The quality of service profile of a particular service or application is defined by the maximum data rate a user can transmit, the guaranteed number of bits provided by UMTS, maximum service data unit (SDU) size, and traffic class. Mainly, as shown in Table 2-1 the traffic class is characterized into four classes according to the delay-sensitivity of the traffic and managed as UMTS traffic demand through RAB [16].

2.3.2 Data Traffic demand in UMTS Network

Once the UMTS network is deployed and start delivering the service, operators need to assess the data traffic demand of their network and upgrade the capacity of the network accordingly to

deliver a better service and satisfy customer need. Most of the time, the usage data traffic collected from the performance measurement reports or call detail record (CDR) is taken as the data traffic demand for capacity dimensioning. However, the total traffic demand offered by the UMTS network is not only the usage (carried) traffic, but also there is blocked traffic which is occurred due to insufficient resources like RAB failure.

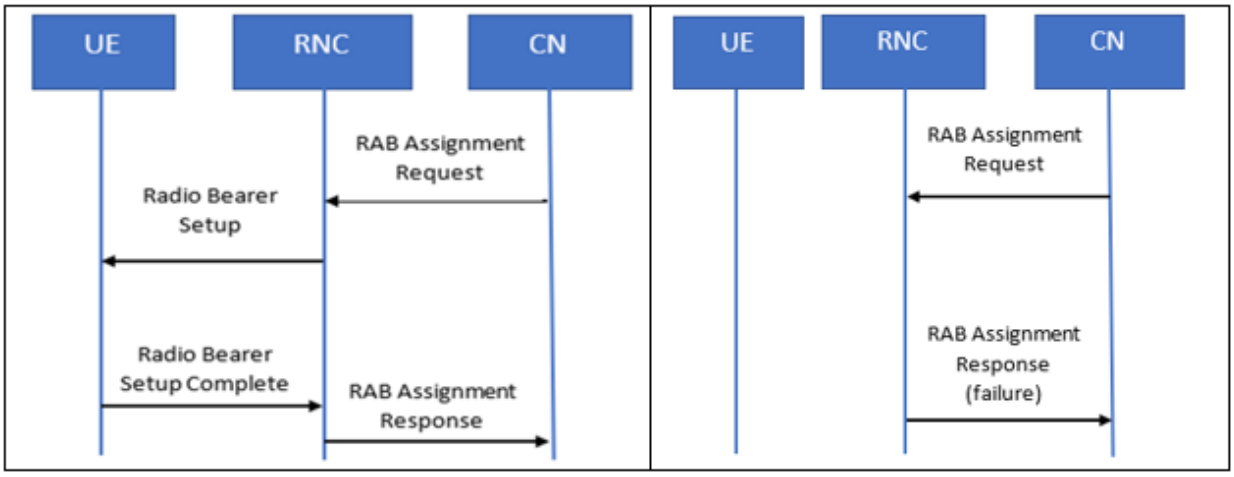


Figure 2-2 Successful and failed RAB Establishment [17].

In the case of the packet service domain, SGSN request RAB service after specifying each parameter of RAB, then the UTRAN establishes selected RAB service depending on the requested QoS, service type, user subscription, etc. [16]. As shown in Figure 2-2, failure to establish RAB connection leads to block the need for data service or interrupt the flow of connection which is occurred due to insufficient resources, and limit the service delivered by the network.

There are some key performance indicators (KPI) used to monitor the RAB services like the number of successful HSDPA RAB establishments for cell and number of failed HSDPA RAB service establishments due to the insufficiency of different resources. These KPIs are also used to monitor RAB status in ethio telecom's UMTS network.

The above-stated KPIs are collected from ethio telecom's performance monitoring tool and depicted in Figure 2-3. In this figure, for instance, on the day of February 16, the number of RAB successes is 100 million whereas the number of RAB failures is 0.5 million. So, the esteemed data traffic demand for the specified period is the sum of success and failures. Moreover, we can also

observe that the average HSDPA RAB failure per day was 1.16 million for the month of February, it is to mean that this figure of demand is lost.

RAB establishment success rate is evaluated every time, and if it is below the threshold the optimization team takes urgent adjustment on the network after analyzing the root cause of the problem. On the other hand, delaying to handle this problem leads to the loss of revenue on the operator side, and also creates customer dissatisfaction.

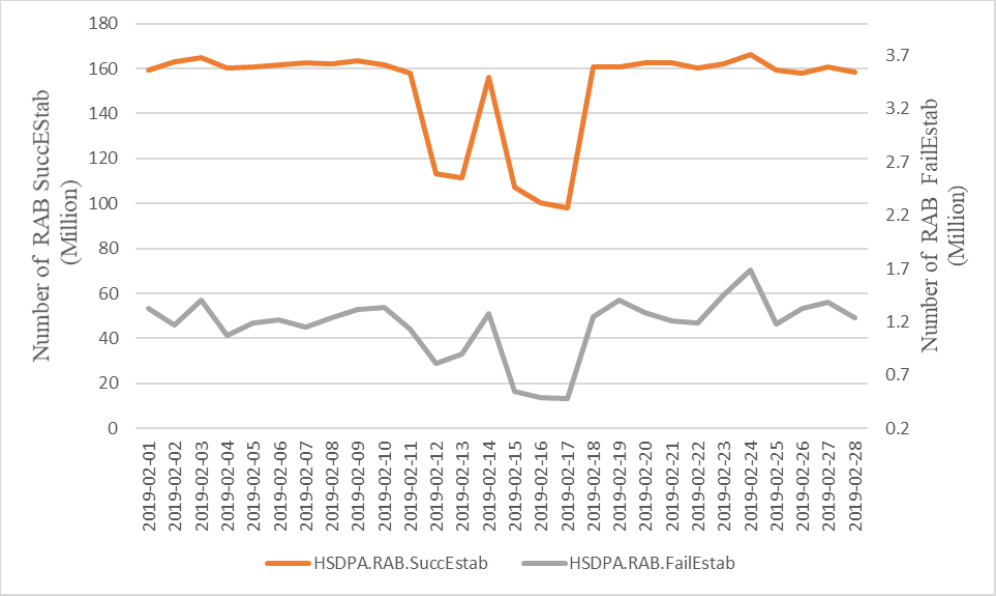


Figure 2-3 Daily HSDPA RAB measurements of UMTS Network.

For this research, the real data traffic demand at a given time and space is defined as the sum of the carried traffic and blocked traffic because of RAB failure, and used for prediction of the future data traffic demand.

The carried traffic (in megabyte) is taken from the downlink data traffic measurement value. But the blocked traffic which is measured in megabyte is not found from the system directly. Hence, it is computed form RAB measurements, which is from the recorded number of RAB failures in the system. Because resources assigned to RAB are useful for root cause analysis in the case of blockage problems for high-speed data transmission [18]. Therefore, the blocked traffic volume is computed from the following formulation assuming that the primary cause of failure is RAB failure due to resource limitation.

$$Bocked\ traffic = DL\ PS\ traffic\left(\frac{\text{RAB.FailEstab}}{RAB.SuccEstab}\right) \quad (2\text{-}1)$$

$$Carrid\ traffic = DL\ PS\ traffic \quad (2\text{-}2)$$

$$Traffic\ demand = Carrid\ traffic + Blocked\ traffic \quad (2\text{-}3)$$

Where,

- RAB.FailEstab – is the number of HSDPA RABs failed to be established due to insufficiency of different resources in the HSDPA serving cell;
- DL PS traffic – is downlink data traffic (in megabyte);
- RAB.SucessEstab – is the number of successful HSDPA RAB establishments for Cell.

3. Mobile Data Traffic and Basics on Deep Neural Network

In this Chapter, an overview of mobile data traffic characteristics from time and space perspectives is presented. Moreover, brief description of deep neural networks, which is used in this thesis to prediction mobile data from in time and space domains, is given.

3.1. Mobile Data Traffic Characteristics and Prediction

Understanding the different characteristics of the mobile data traffic and forecasting future demand, accordingly, helps for the effective planning of mobile networks and efficient utilization of network resources during the operation phase. For instance, incorporating the forecasted traffic is necessary during energy-efficient operation of cellular networks using cell zooming (i.e. regulate cell size according to network load) [19]. Additionally, from the mobile users' perspective, understanding the congestion (traffic) level of the network using some mechanisms has a great benefit as users can select an appropriate time and/or space while entertaining service with a minimum cost [9]. To this end, to select the most appropriate prediction approach it is better first to understand the basic characteristics of mobile data traffic; and it is presented in the following sub section.

3.1.1 Mobile Data Traffic Characteristics

To manage resources using data pricing, energy-saving, and congestion control mechanisms, the observation of temporal and spatial characteristics is more important. Mobile data traffic has different characteristics that can be expressed from temporal and spatial aspects. The foremost temporal data traffic characteristics are periodicity and seasonality. In periodicity, the data traffic is repeated with a certain frequency (e.g. hourly). But, in seasonality, the pattern is repeated at specific regular intervals less than a year (e.g. daily, weekly, monthly, quarterly) [4]. Additionally, the trend of data traffic is also shown the propensity of the traffic to decrease or increase over time.

For instance, as shown in Figure 3-1, which is a one-week mobile data traffic usage collected from Addis Ababa city's UMTS network. The traffic follows the same usage pattern each day of the week; it reaches to its peak at night times around 10:00 PM and reaches its minimum value in the morning around 4:00 AM.

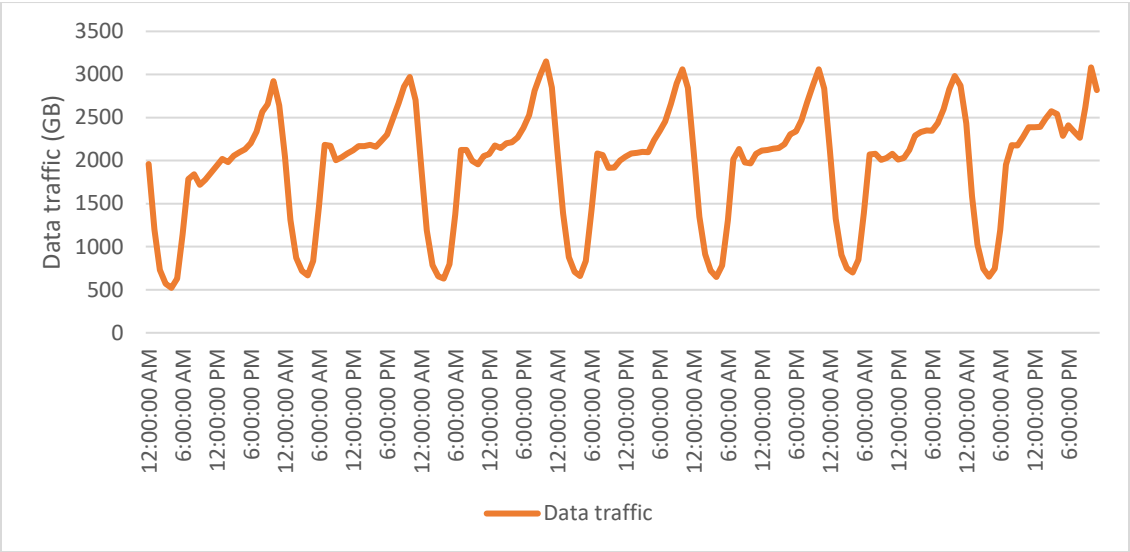


Figure 3-1 One-week mobile data traffic for Addis Ababa city's network.

This figure provides the time series characteristics than spatial characteristics since the data is aggregated for the specific granularity of the city's network. However, the data consists the spatial concept, it doesn't show the spatial distribution of the data traffic in the city.

Geographical hierarchy and distance are the primary characteristics of the spatial aspect of data traffic. In geographical hierarch, for example, the data traffic collected from sites level, also belongs to the area were the sites are found. The distance between the two areas can measure the correlation between the two areas, for instance, the adjacent area is more similar than distant one [4].

The spatial distribution of mobile data traffic for Addis Ababa city's network is shown in Figure 3-2. In this figure, the inhomogeneity of mobile data traffic distribution is observed with different colors in the city.

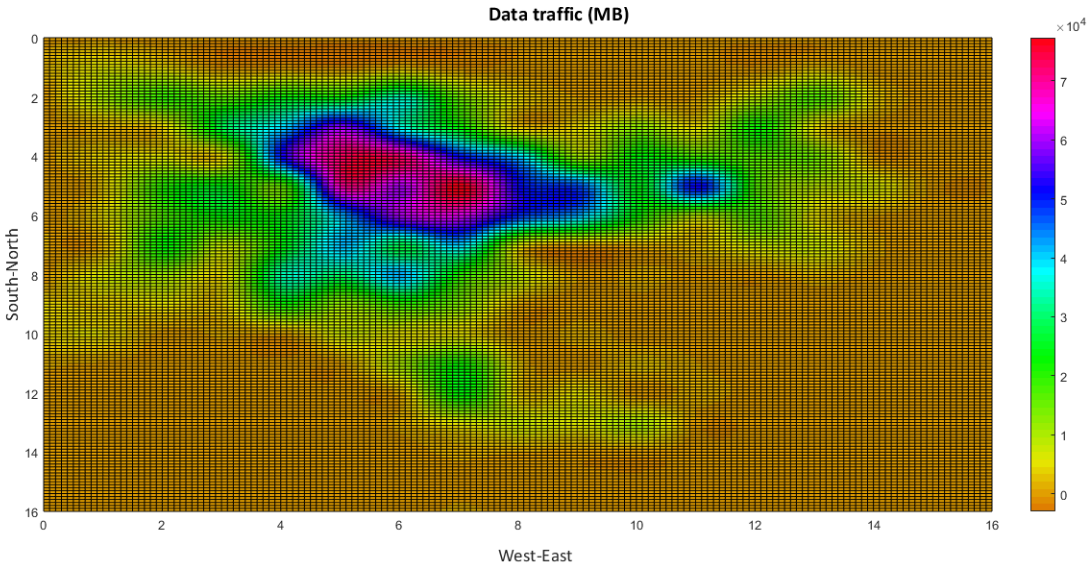


Figure 3-2 Spatial data traffic distribution for Addis Ababa city's network at 4:00 PM

The spatiotemporal distribution of mobile data traffic for Addis Ababa city's network is illustrated in Figure 3-2 and Figure 3-3. We see that the traffic varies in different areas, possibly due to the nature of users' mobility. In Figure 3-2, the deep red color indicates the area with high data traffic. At the day time, i.e., around 4:00 PM, the data traffic demand increases towards the center of the city, which are business areas, whereas in Figure 3-3, during night time, i.e., at 12:00 AM, it starts to move towards the border of the city, to the residential areas.

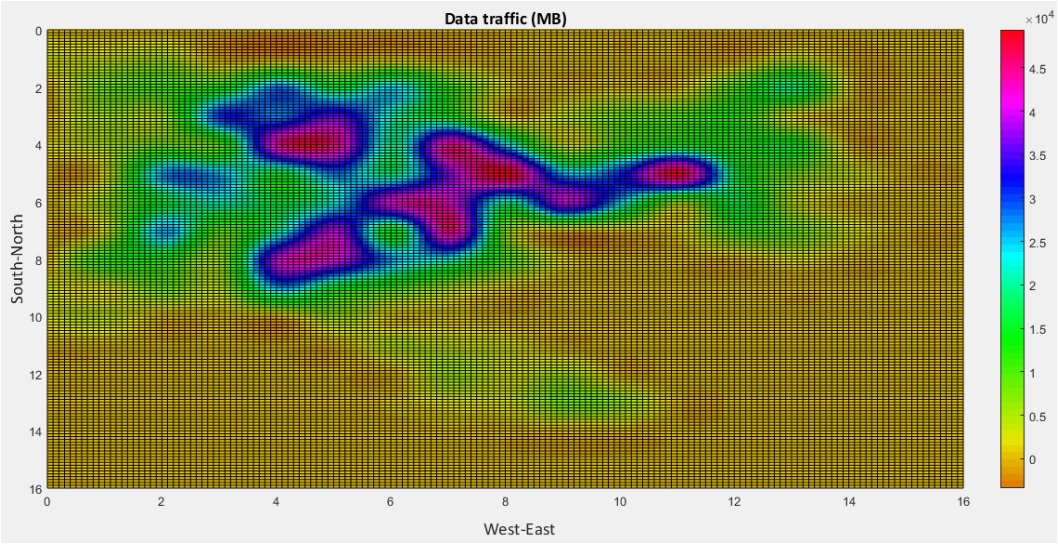


Figure 3-3 Spatial data traffic distribution for Addis Ababa city's network at 12:00 AM

Furthermore, the characteristics of spatiotemporal mobile data traffic are dependent on geographical locations, population, mobility of users, and other factors. In addition to this, a social event like a national holiday, a spiritual ceremony or soccer game has an impact on it [12]. Besides, a proper prediction model is needed, which can capture this dynamic behavior of the data traffic and make the prediction with better accuracy.

3.1.2 Mobile Data Traffic Prediction

Predicting mobile data traffic demand is one of the main tasks of operators to accommodate the future increasing data traffic demand. Mobile data traffic prediction is studied from a temporal, spatial, and spatiotemporal perspective using different approaches. Then, the basics regarding the above-mentioned predictions are presented below.

a) Temporal Data traffic Prediction

Temporal data traffic prediction is an approach that effectively exploits the temporal characteristics of traffic patterns from historical observations to infer relevant information about the future. The network traffic prediction methods are described under four categories as follows [20]

i. Linear time Series techniques: - are covariance structure in time series. Auto Regressive (AR) and the Moving Average (MA) models are the two main sub groups of this technique. These two models are combined to generate the auto regressive moving average models, for instance, ARMA and SARIMA models.
ii. Nonlinear time series techniques: - are used to explore features that can't be handled by linear processes such as cycle and time-change variance. Example, Neural Network techniques.
iii. Hybrid model techniques: - It is mainly the combination of linear and nonlinear models. Example, the combination of ARIMA and Multilayer Artificial Neural Network (MLANN)
iv. Decomposed model: - is implemented by decompose the time series in to seasonal, trend, cyclically, and Irregular components. For example, the nonlinear decomposed

model implemented by decomposed time series into trend, period, mutation component and a random component.

b) Spatial Data Traffic Prediction

In spatial prediction, the mobile data traffic patterns are analyzed by identifying spatial environment factors, such as topology and user mobility, in order to make statements of the likelihood for an event with spatially correlating occurrences. For instance, the hidden spatial traffic pattern between base stations is investigated and applied to select the very important base stations. These base stations' data traffic represents the characteristics of the entire network. Then, by applying their data traffic, the entire system traffic is predicted [21].

c) Spatiotemporal Data Traffic 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. Even though the joint consideration most likely improves the prediction accuracy, technically spatiotemporal dependencies in cellular networks are complicated due to the following reasons [12]:

- Mobility of users across spatially distributed base stations;
- Geographical distribution of mobile traffic, which is influenced by social activities, population, and holidays;
- User behavior and diverse network demand for internet-based applications;
- Heterogeneous network architecture.

To this end, deep neural network models are widely used to capture its complex nature of the temporal dynamics and spatial correlation of data traffic [11] [8] .

3.2. Basics on Deep Neural Network

The idea of Artificial Intelligence (AI) is started around 1956, which deals with enabling machines to impersonate as human beings. One of the enablers of AI is machine learning, which comprises

the techniques that empower computers to identify things from given information and provide AI applications. In the meantime, deep learning, which is a subset of machine learning, is more applicable for the data analysis with increased nonlinearity and more complex in feature extraction, as shown in Figure 3-4 [22].

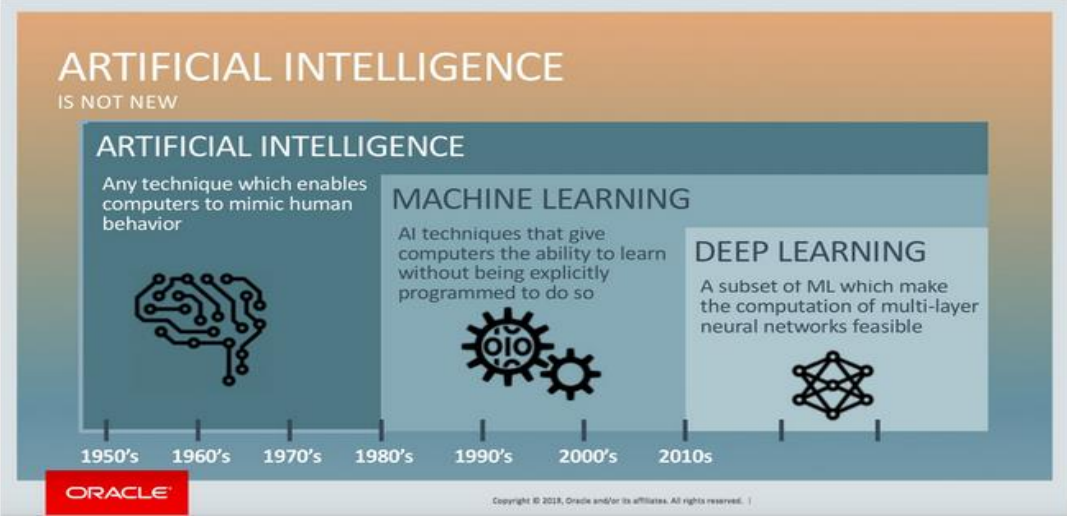


Figure 3-4 Artificial intelligence, Machine learning, Deep learning [22].

The structure of the human brain is taken as a means to generate the idea of deep learning. Neural network architecture is applied by most deep learning methods; hence it is referred as a deep neural network. The word "deep" denotes the number of hidden layers on the neural network [23]. The architecture of the neural network mainly has an input layer, hidden layer, and output layer, as shown in Figure 3-5. The input and output layers consist of input neuron and output neuron, respectively, and in the middle of them found hidden layers. The hidden layer's input and output are not visible because the output of one layer is already taken as input for the other layer.

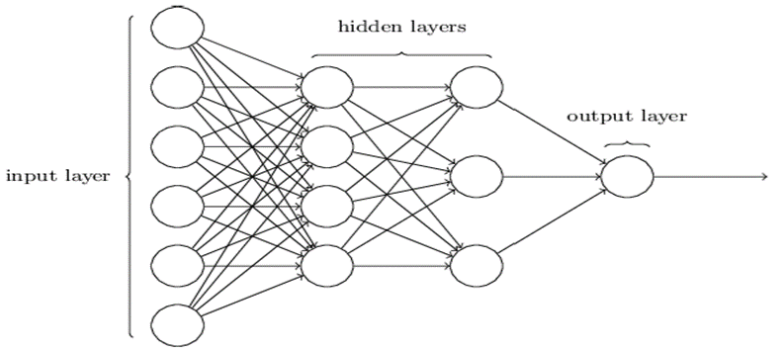


Figure 3-5 Input, output and hidden layers of Neural Network [24]

In order to understand the implementation of deep neural networks, it is essential to provide a brief overview of the following concepts.

a) Forward Propagation

Forward propagation is the process of feeding the output of each layer to the next layer as input after passing it through a mathematical function named activation function. The activation function mainly adds non-linearity to the network, and the commonly used functions are hyperbolic tangent (*tanh*), rectified linier unit (*ReLU*), sigmoid, and *Softmax* [25].

b) Backward Propagation

During the training process, once the forward propagation is done, the output is measured using loss functions like mean absolute error. The loss function measures the difference between the target and the final output. Then after, parameters, set of biases and weights are readjusted (learned) to minimize the error. In doing so, some tunings are done in these parameters using backpropagation algorithms based on gradient descent.

Optimizing with gradient descent iteratively attempts to find the minima of a function by taking small steps towards the minimum value of the function. There are different gradient-based optimization functions, such as stochastic gradient descent (SGD) and adaptive momentum estimation (Adam) [26]. In a deep neural network, the optimization process gets difficult when the network becomes complex because of exploding or vanishing gradient problems. Therefore, the models are required to handle these problems.

c) Exploding and vanishing gradient

In the backward propagation process, as we go backward, there may be a too large gradient or too small gradient, exploding and vanishing gradient, respectively. Exploding gradient creates instability in the model learning process. Whereas in vanishing gradient, the earlier layers don't get a new update and the model encounter difficulties to optimize these layers. This kind of problem happens when there is a need for investigation on long-term dependencies, which is the link between the current and the earlier information provided for the desired prediction [27].

3.3. Deep Neural Network on Mobile Data traffic Prediction

In mobile network traffic prediction deep neural network plays a big role as it is capable of extracting high-level features from the data that has inner correlation and complex structure. Moreover, it handles large amounts of data, multi-task learning, and geometric mobile data learning which refers to multivariate data represented by order, topology, metrics and coordinates [28].

Various deep learning models applied for spatiotemporal mobile data traffic prediction in different approaches such as CNN, RNN,3D CNN, ConvLSTM and GNN. Moreover, the combination of CNN and RNN is widely used on data traffic prediction and contribute more improvement in prediction performance accuracy [11, 8, 12]. The basic concepts of these models are discussed below.

3.3.1 Convolutional Neural Network

CNN is a type of deep neural network initially designed for image processing problems, but now it is applied to different fields of studies like speech recognition, natural language processing, and time-series data prediction. CNN is organized by different layers that have specific tasks. Besides the input layer and output layer there are three stacked layers; namely, a convolutional layer, pooling layer and fully connected layer, as shown in Figure 3-6.

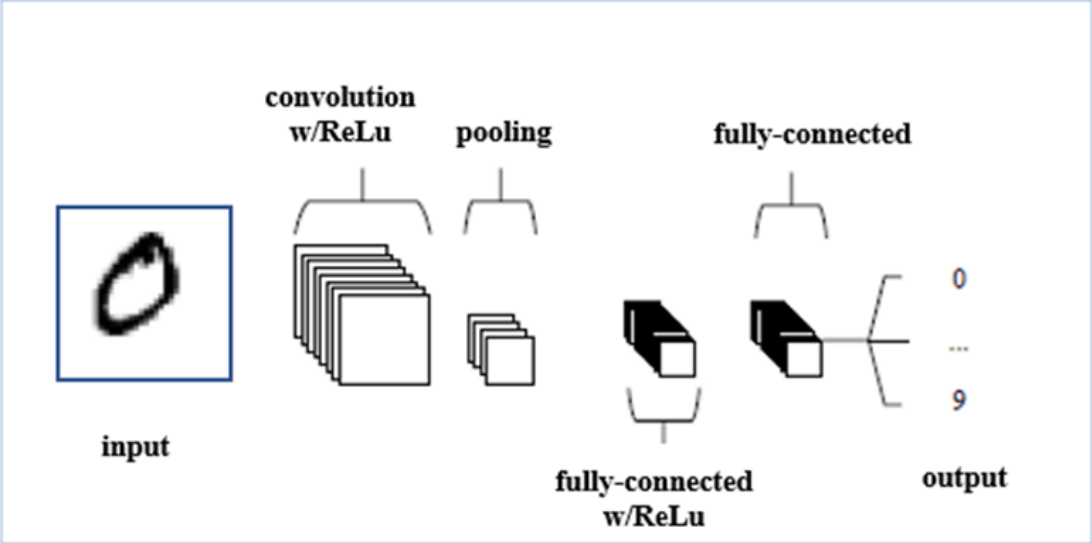


Figure 3-6 A simple CNN architecture [29].

The convolutional layer is used to extract significant features from the pixel values of the image held by the input layer, its parameters focused on the use of kernels. The kernels, also known as convolution filters, are implemented as matrices and applied to the small area of the input layer to calculate the dot product between them and move to the entire input layer to get the intended feature. The pooling layer down-samples the spatial dimensionality of the input according to the pooling policies like max-pooling or mean-polling. Finally, the fully connected layer flattens the result as of regular neural network and then produces the desired output whether classification or prediction [29].

CNN can be implemented to exploit multi-dimensional data such as one-dimensional CNN for speech processing, 2D CNN for image classification, and 3D CNN for object processing [30].

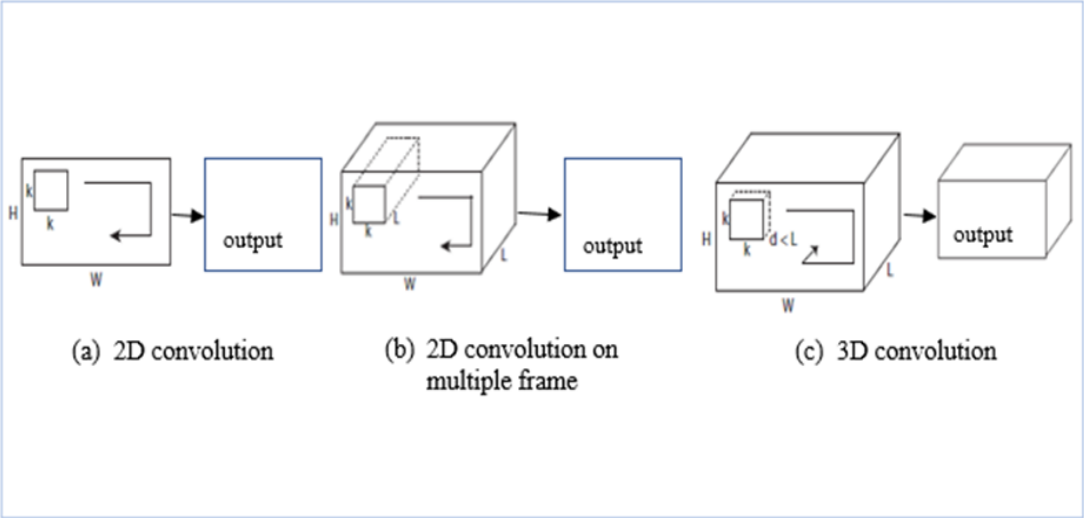


Figure 3-7 2D and 3D convolutional operations [31].

2D CNN and 3D CNN are more utilized for different studies because of their capacities for capturing the spatial complexity of the desired object. Moreover, 3D CNN has capabilities of modeling space and time information whereas the temporal aspect of the input object diminishes after the first convolution in 2D CNN, but for multiple images, it can handle them as different channels [31]. Figure 3-7 shows the input and output convolutional operations of 2D and 3D CNN and discussed as follows:

a) Appling 2D convolution on image results in an image;
b) Appling 2D convolution on video volume (multiple frames as multiple channels) also results in an image;
c) Appling 3D convolution on video results in another volume, preserve temporal information of input signal.

CNN has a lot of advantages to deal with a multidimensional dataset, mainly it has the capability to extract spatial relationships from the dataset using its kernel.

3.3.2 Recurrent Neural Network

In the learning process of feedforward neural network, the data flow in one direction from the input to the output layers, so that it is difficult to recognize previous patterns of the dataset on

the process of predicting sequential data. This kind of problem is solved by RNN which has a feedback loop with its repeated modules on the network [32]. However, the main limitation of RNN is unable to learn long-term dependencies due to vanishing gradient problem in the operation of backpropagation.

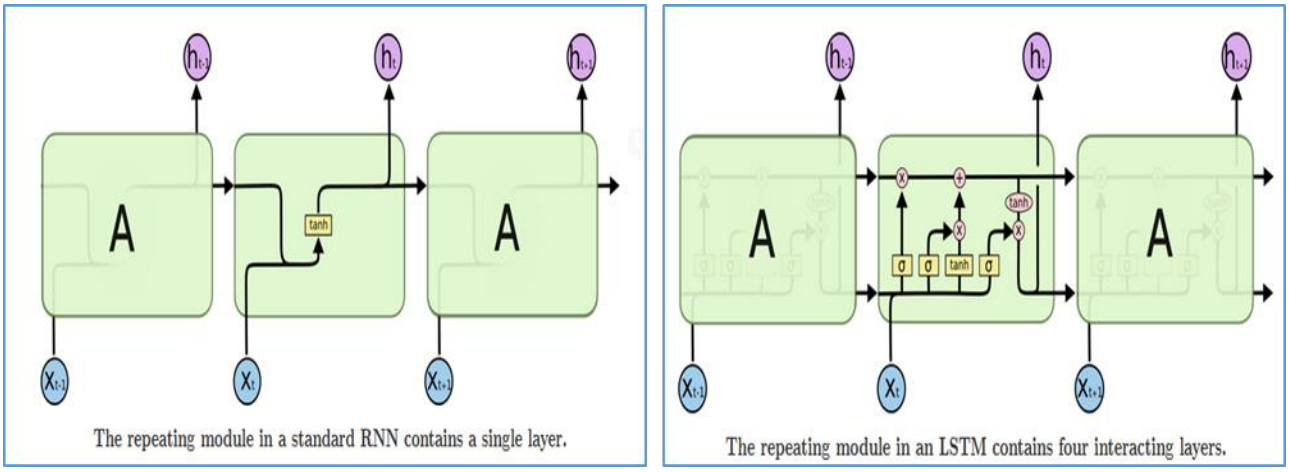


Figure 3-8 Standard RNN and LSTM [33].

One of the solutions used to overcome the vanishing gradient problem is LSTM, which is a special kind of RNN. LSTM handles vanishing gradient problems easily by using its four interacting layers within a cell which differs from standard RNN models as shown in Figure 3-8. This layer can help LSTM to learn long term dependencies easily and also it helps to improve the performance of the model by implementing three gates as follows below [34].

- ***forget gate***: - which helps to select only the necessary information from the previous state
- ***input gate***: - which decides to let go new information based on its significance in the current step before the current state is updated.
- **outp*ut gate*: -** which will decide which information of the cell is passed to the output.

The formulation to implement this process is presented below [35].

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (3\text{-}1)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}f_{t-1} + W_{cf}c_{t-1} + b_f) \quad (3\text{-}2)$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc} x_t + W_{hc} h_{t-1} + b_c) \quad (3\text{-}3)$$

$$o_t = \sigma(W_{xo} x_t + W_{ho} h_{t-1} + W_{co} c_t + b_o) \quad (3\text{-}4)$$

$$h_t = o_t \tanh(c_t) \quad (3\text{-}5)$$

Where *i, f, o, b, h* and *c* are the input gate, forget gate, output gate, bias, hidden state, and cell respectively, and σ is the logistic sigmoid function. The weight matrix subscripts, W_{hi} is the hidden-input gate matrix, W_{xo} is the input-output gate matrix, etc.

LSTM is a powerful model for predicting sequential data like the time series data; it has the capability of capturing long-term dependencies of the dataset.

3.3.3 Convolutional Long Short-Term Memory

ConvLSTM is a special type of LSTM network which incorporating convolutional operator in each LSTM cells. The convolutional operation helps to capture the spatial properties of the data, in addition to learning the long-term dependencies, which are the unique behavior of LSTM. ConvLSTM first applied on nowcasting prediction problem, convolution operators have replaced the matrix products within LSTM cells, so that, the model can handle the reading of two-dimensional data as row and column easily and work on the spatiotemporal aspect of the dataset by observing their time dependencies. The main formulation of ConvLSTM is described below [36].

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} O C_{t-1} + b_i) \quad (3\text{-}6)$$

$$f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} O C_{t-1} + b_f) \quad (3\text{-}7)$$

$$C_t = f_t O C_{t-1} + i_t O \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \quad (3\text{-}8)$$

$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} O C_{t-1} + b_o) \quad (3\text{-}9)$$

$$H_t = o_t O \tanh(c_t) \quad (3\text{-}10)$$

Where cell outputs C_t, inputs X_t, hidden states H_t, and gates i_t, f_t, o_t of the ConvLSTM are 3D tensors, its last two dimensions are spatial dimensions (rows and columns), and '$*$' denotes the convolution operator and 'O' denotes the Hadamard product:

ConvLSTM is well-suited for spatiotemporal data investigation. It is applied in different fields of studies such as video frame forecasting and it is adopted for spatiotemporal data traffic prediction in [11]. Hence, the model is also applied in this research, since it is capable of handling spatial correlations and temporal dynamics of the dataset.

4. Results and Discussion

In this Chapter, the mobile data traffic dataset is described from the temporal and spatial aspects. In addition, the model and tools used for the experiment are also well defined. The prediction approach is optimized after the training is done using different hyperparameters, and the results are discussed according to the nature of the mobile data traffic.

4.1. Proposed approach

The prediction approach applied here is used from Keras, which is a deep learning library [37]. The overall process passes through three phases as shown in Figure 4-1. At the input phase shown at the bottom of the figure, the processed data is reshaped and is passed to the model. Then, the model observes twenty-four-hour data and predicts the next six-hour data traffic of two features (Carried traffic and Blocked traffic). Finally, by summing up these two predicted data traffic, it will provide real data traffic demand at a given space and time.

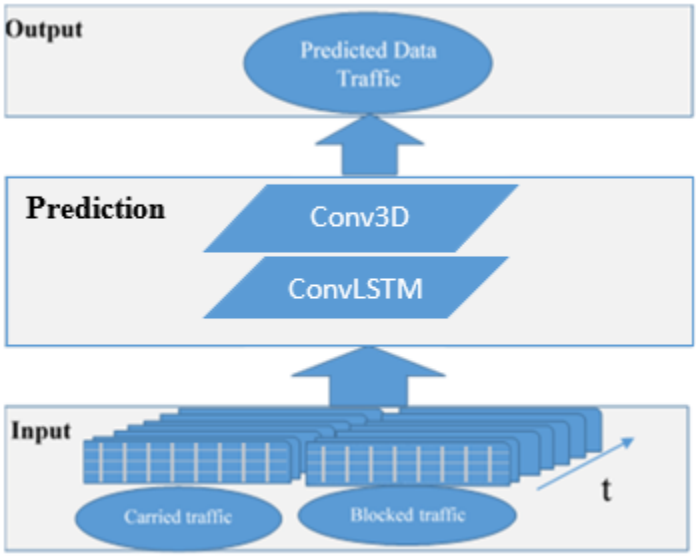


Figure 4-1 Mobile Data traffic prediction approach.

In the prediction phase, the layers are stacked by ConvLSTM models staring from the input layer, and at last three-dimensional convolution (Conv3D) is applied to handle the output layer. ConvLSTM model is preferred for the implementation of spatiotemporal data traffic prediction

because of the spatial and temporal nature of the dataset. Its convolutional operation helps to capture the spatial dependencies and the LSTM captures the temporal dynamics of the data traffic.

To improve the performance of deep learning models, one of the preferable methods is batch normalization, which performs normalization on the activation of previous layer at each batch. So, it is applied here between each layer transition [38]. For the implementation of the model open source libraries Keras and Tensorflow [39] are used in the Python environment and tested on a high-performance computer.

4.2. Dataset Preparation

The dataset used for this research is collected from ethio telecom's performance data measurement report from January 30 to April 30, 2019 on the hourly basis. It has mainly the records on downlink data traffic volume, which is taken as carried traffic. In addition to this, the number of successful HSDPA RAB establishments and the number of failed HSDPA RAB establishments are collected; which are used to compute the blocked data traffic as described in Chapter2. Furthermore, the site's location coordinates (longitude and latitude) are collected. The data processing part follows steps that are explained in the following steps.

First, the area is selected from the Addis Ababa city ethio telecom's network service area, which covered 49 Km^2 as shown in Figure 4-2 (a). It is covered with densely distributed sites that have more spatiotemporal dependencies introduced by user mobility, population, and social activities.

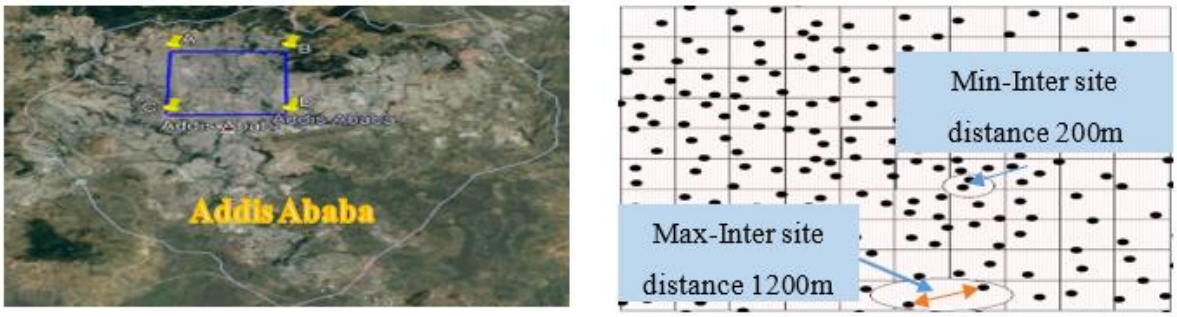
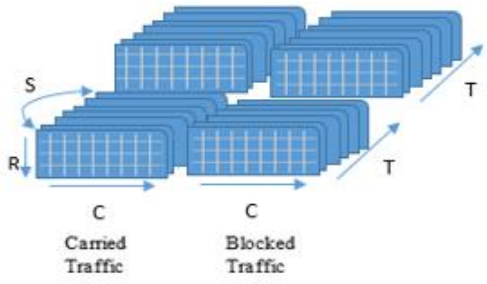


Figure 4-2 (a) Target area (b) Spatial distribution of sites on the target area

Then, the geographical grid is prepared for the selected area to get spatial information on data traffic demand. The area is divided into 100 grids which have 700*700 m^2 grid resolution, and then aggregate data traffic per each grid is collected. In the selected area the maximum inter-site distance is 1.2km and the minimum inter-site distance is 200m as shown in Figure 4-2 (b). So, the grid resolution is set to be 700*700 m^2 which is the average of both distances.

Finally, the dataset is processed for 100 grids and arranged sequentially. Ninety percent of the dataset is used for training and validation of the model, and the remaining ten percent is used for testing. To represent the spatial information and temporal aspects, the dataset is reshaped to five-dimensions. As shown in Figure 4-3, row (R) and column (C) represent spatial information of the dataset, sample (S) represents the number of instances used for prediction, time steps (T) represents the number of hours per sample. Carried and blocked traffic are the features to be predicted.



Where: - S= number of samples

R= row

C= Column

T= number of hours per sample

Figure 4-3 Reshaped input dataset

4.3. Model Performance Evaluation Metrics

The evaluation is performed using three standard prediction evaluation metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE), which help to discover the error characteristics from different aspects [40]. RMSE can help to figure out high-level error values, and MAE help to get the average error within a specified range of the predicted dataset. The values of MAPE and their interpretation, which are shown in Table 4-1 are used for the explanation of the prediction accuracy.

Table 4-1 Interpretation of MAPE value [41]

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

The formulas for evaluation metrics are given in the below equations.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \quad (4\text{-}1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \quad (4\text{-}2)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{Y}_i - Y_i|}{y_i} \times 100 \quad (4\text{-}3)$$

where ***n*** is the number of instances and the difference between the target value (Y_i) and the predicted value ($\hat{Y}_i$) is defined as e_i.

4.4. Hyperparameters Tuning

Hyperparameters are external parameters whose values are used to control the learning process. Here ConvLSTM model is explored with different hyperparameters. Selecting these appropriate combinations of hyperparameters is critical to learn and predict the desired task with better accuracy. The hyperparameters which have a significant impact on the training of this model and which can be modified according to the desired task are selected and prepared. The list of hyperparameters and their values is depicted in Table 4-2.

Table 4-2 Hyperparameters value

Hyperparameters	Value
Filters	32,64
Kernel size	(3,3), (5,5)
Layers	3,4
Learning rate	0.001,0.002,0.005,0.01

Filters can detect more features that help to extract different data traffic usage behavior of the network. So, an increasing number of filters from 32 to 64 improve the prediction performance of the model, by contributing additional information for the training process. Adding more layers is also helping to detect abstract information from the available dataset. But, in this case, the model with three ConvLSTM layers has less error than the model with four layers, as shown in Figure 4-4. So, applying more layers does not have a significant impact on the performance of the model.

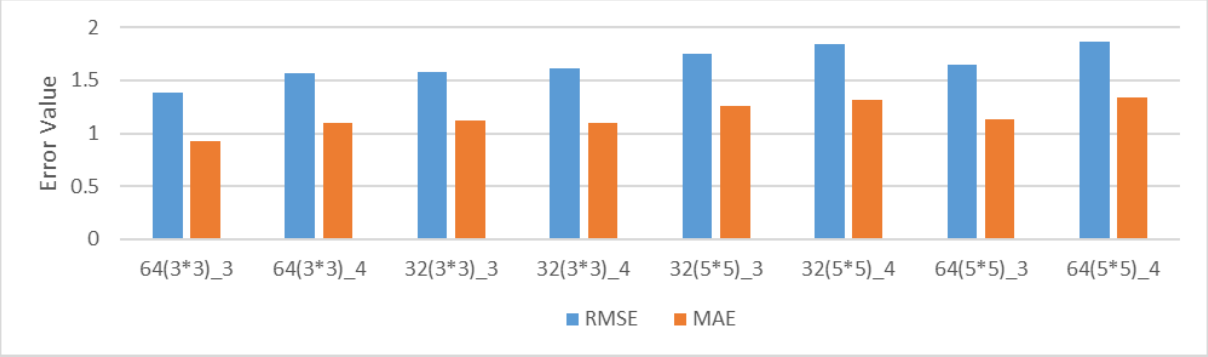


Figure 4-4 RMSE and MAE value of Carried traffic for ConvLSTM model

The height and width of the convolution window are defined by kernel size, and mostly increasing the kernel size helps to see additional information within specified grids. However, in this case, the model with kernel size 3*3 has better performance than kernel size 5*5 which has an RMSE value above 1.6 as shown in Figure 4-4. This is because of the spatial characteristics of the dataset where it is not providing the relevant information to the specified grid as the kernel size increased.

For instance, as shown in Figure 4-5, the spatial correlation between grid 25 and its neighbor grids at 3*3 kernel size (green square) is above 0.8 for all, but in 5*5 kernel size (yellow square) correlation values of nine grids are below 0.8, which indicates their loose correlation with grid 25.

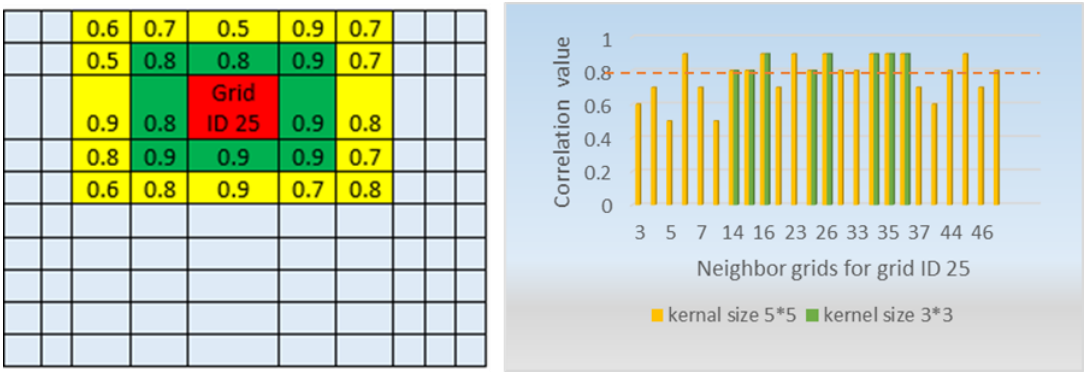


Figure 4-5 Spatial correlation between grid ID 25 and its neighbor grids

From the assessment, it shows that the correlation of data traffic between the grids found in 2.1km^2 (3*3 kernel size) is high and provide more information for prediction performance improvement, but the information above 2.1km^2 doesn't have significant contribution, because the sites which are far beyond 2km have less correlation with the site found in specified grid as shown in Figure 4-5, so applying 3*3 kernel size is more preferable than 5*5 kernel size for this task.

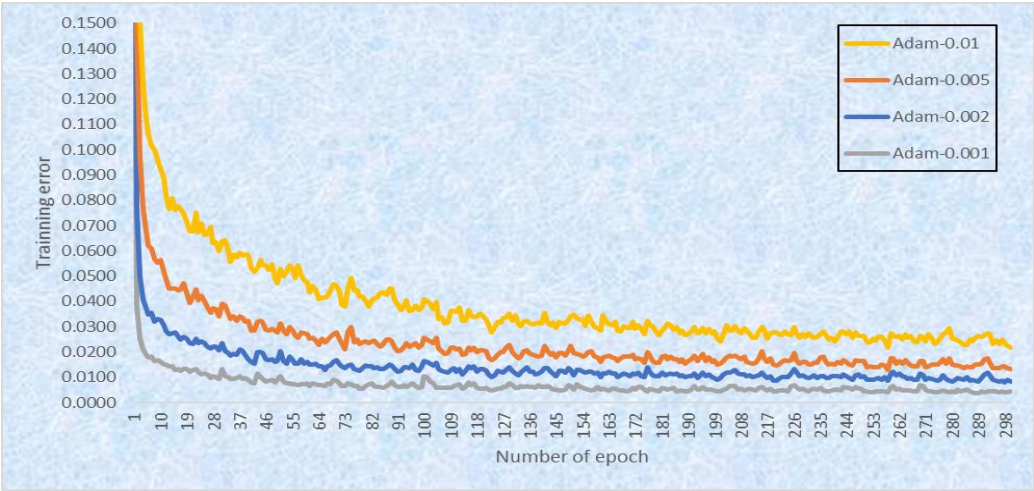


Figure 4-6 Adam optimizer with different learning rate

Optimization function and its learning rate have a significant role in the improvement of the prediction performance of deep learning models. Studies disclosed *Adam* is a more powerful optimizer for practical application than another stochastic optimization method [42]. So, in this training process, *Adam* optimizer is applied and tested using different learning rates. The learning rate is used to control how quickly a model learns the problem. As shown in Figure 4-6, with the learning rate 0.002 the learning process is stable and optimal, so the model is able to learn the problem better than the others.

To this end, each prediction approach is tested using eight instances from the test dataset, and the worst prediction performance of the model according to the evaluation metrics from these eight tests (instances) are taken for the comparison. In Table 4-3, the model comparisons based on hyperparameters, and the impact of the changes in the value of hyperparameters on the model's performance are described. In this table, ConvLSTM with three layers, 64 filters, 3*3 kernel size, and one Conv3D model (ConvLSTM(3*3)_64(3*3)_64(3*3)_Conv3D) perform better prediction than the others.

Table 4-3 RMSE and MAE values for Carried and Blocked traffic

Model	Carried traffic		Blocked traffic	
	RMSE	MAE	RMSE	MAE
ConvLSTM(3*3)_64(3*3)_ 64(3*3)_Conv3D	1.3701	0.8458	0.0022	0.00127
ConvLSTM(3*3)_64(3*3)_64(3*3)_64(3*3)_Conv3D	1.5636	1.1001	0.0039	0.00168
ConvLSTM(3*3)_32(3*3)_ 32(3*3)_Conv3D	1.5838	1.1169	0.0035	0.00216
ConvLSTM(3*3)_32(3*3)_32(3*3)_ 32(3*3)_Conv3D	1.6172	1.0981	0.0018	0.00104
ConvLSTM(5*5)_32(5*5)_ 32(5*5)_Conv3D	1.7511	1.2597	0.0036	0.00221
ConvLSTM(5*5)_32(5*5)_32(5*5)_ 32(5*5)_Conv3D	1.8453	1.2878	0.0027	0.00144
ConvLSTM(5*5)_64(5*5)_ 64(5*5)_Conv3D	1.6462	1.1349	0.0022	0.00132
ConvLSTM(5*5)_64(5*5)_64(5*5)_ 64(5*5)_Conv3D	1.8677	1.3428	0.0024	0.00138

4.5. Observation on Prediction Results

Providing the twenty-four-hour historical data, the prediction is made for the next six-hour data traffic demand. As shown in Figure 4-7 hourly time-series trends and predicted data traffic for one grid, the selected approach (ConvLSTM with 64 filters, 3*3 kernel, and three layers) learn the data traffic demand history and predict closer values to the target than the others. The accuracy decreases when the prediction step goes from the first to the last because of error propagation from the previous predictions.

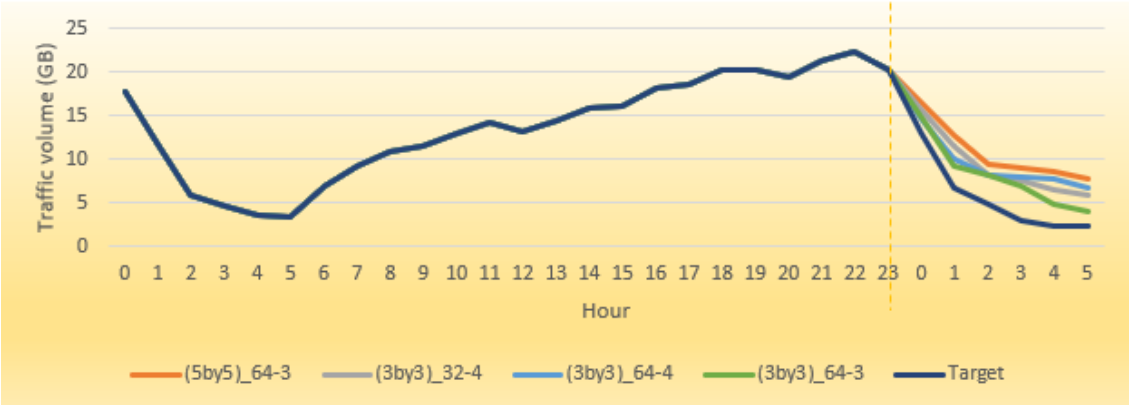


Figure 4-7 Time series trend and predicted data traffic for one grid

Observing the prediction from the spatial point of view, as shown in Figure 4-8, the selected approach (ConvLSTM(3*3)_64-3) can capture the intensity of the data traffic in each grid closer than the others. For instance, at T+1, where T represents a twenty-four-hour data traffic history, the target data traffic at grid id 22 is 8.5 GB, and for the selected model prediction is 8.9 GB which shows a 0.4GB difference. Whereas the other model is predicted as 6.1GB which shows about a 2GB difference. So, the selected model predicted more accurate than the other models.

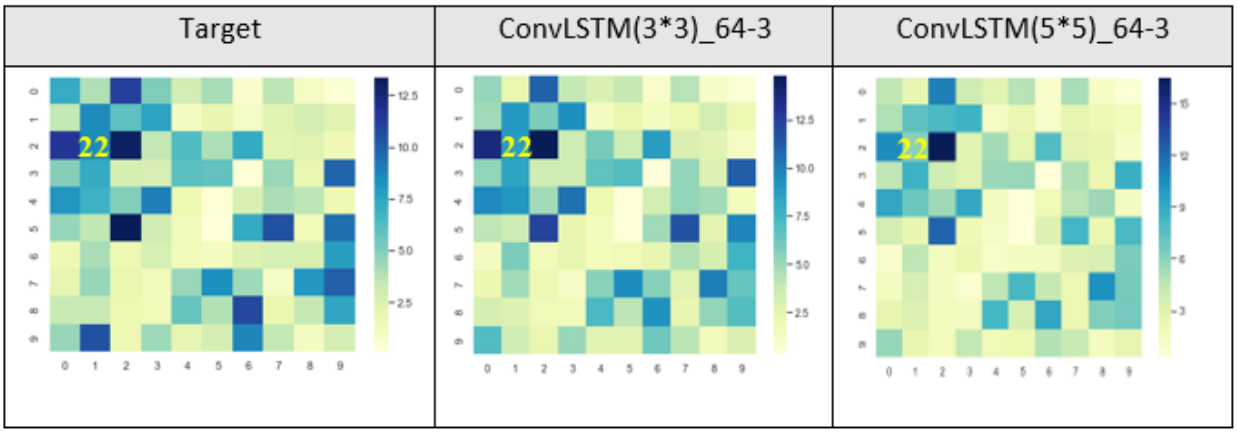


Figure 4-8 Spatial views of target and predicted data traffic at T+1

As shown in Figure 4-9, At T+1 out of the total 100 grids 31 grids are predicted with a high accuracy, and 28 grids are predicted with good accuracy (with the error margin of 10 to 20 percent), 30 grids are predicted with reasonable value, while 11 grids are predicted inaccurately with the MAPE value of above 50%.

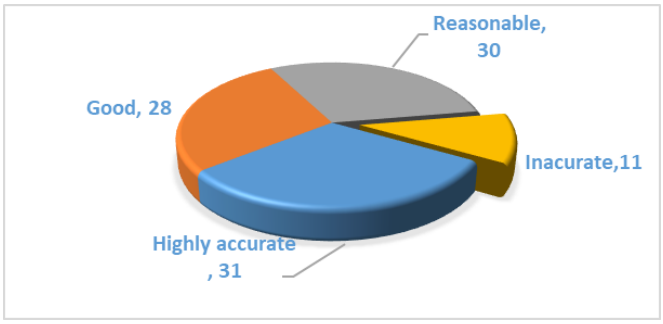


Figure 4-9 MAPE results for 100 grids

The eight grids from eleven grids that have high MAPE value, the inter-site distance between their neighbor grids is above 800m, and four of them are found at the edge of the target area. So, the information obtain from neighbor grids decreased due to the loose correlation between grids, which have an impact on the degradation of prediction performance.

4.5.1 Impact of Grid Size on Prediction Performance

In this section, the comparison is presented for two scenarios in the selected area, which are differentiated with the number of grids but applying the same model for two of them.

Scenario 1: grid resolution with 700*700 m^2, which is already preprocessed in Section 4.2

Scenario 2: grid resolution with 350*350 m^2, which is described in Figure 4-10

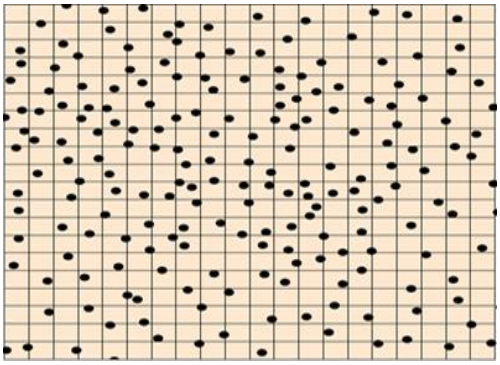


*Figure 4-10 Spatial distribution of sites for 350*350 m^2 grid resolution*

The comparison of prediction accuracy is performed for the selected 39 grids, which are found in both scenarios with the same site's distribution. Moreover, as shown in Figure 4-11, Scenario 1 has a better prediction performance than scenario two. For Scenario 1, thirty-three grids are above reasonable MAPE value, which has less error-prone than the other one, whereas, in scenario two, twelve grids are predicted inaccurately out of 39 grids.

Table 4-4 Summary of grids for two scenarios

Scenario	Grid Resolution	Number of grids with site	Number of grids without site	Total grid
1	700*700 m^2	100	0	100
2	350*350 m^2	182	218	400

In Scenario 2, the grids that have site is less than the grids without sites (218), as shown in Table 4-4. Implementing the model with more zero value grids decreases the prediction performance because the information originates from zero value grids has no significant contribution to the learning process.

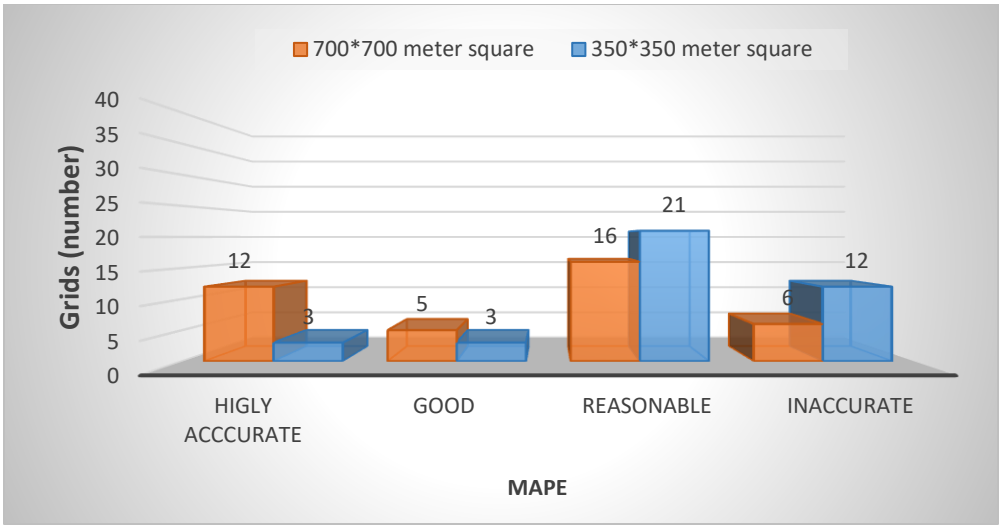


Figure 4-11 MAPE comparison result for different grid resolution

Predicting the data traffic demand at a small grid resolution helps the operators to know the customer's data traffic needs more closely, but applying the data traffic collected in the site level for predicting data traffic demand in a small grid size is not applicable. In the meantime, preparation on data traffic demand for small grid resolution from site-level traffic measurement needs to do further study.

4.5.2 Observation on Predicted Data Traffic Demand with Blocked Traffic

In addition to predicting the real data traffic demand, investigating the predicted blocked traffic at the same time and space gives a significant insight into the network maintenance and optimization decision. As shown in Figure 4-12, carried traffic at specified day is increased (at point A), and at the same time, the blocked traffic also increased at this area (at point B). Whereas the total data traffic demand, which is the sum of carried traffic and blocked traffic at some neighbor grids is low (at point C).

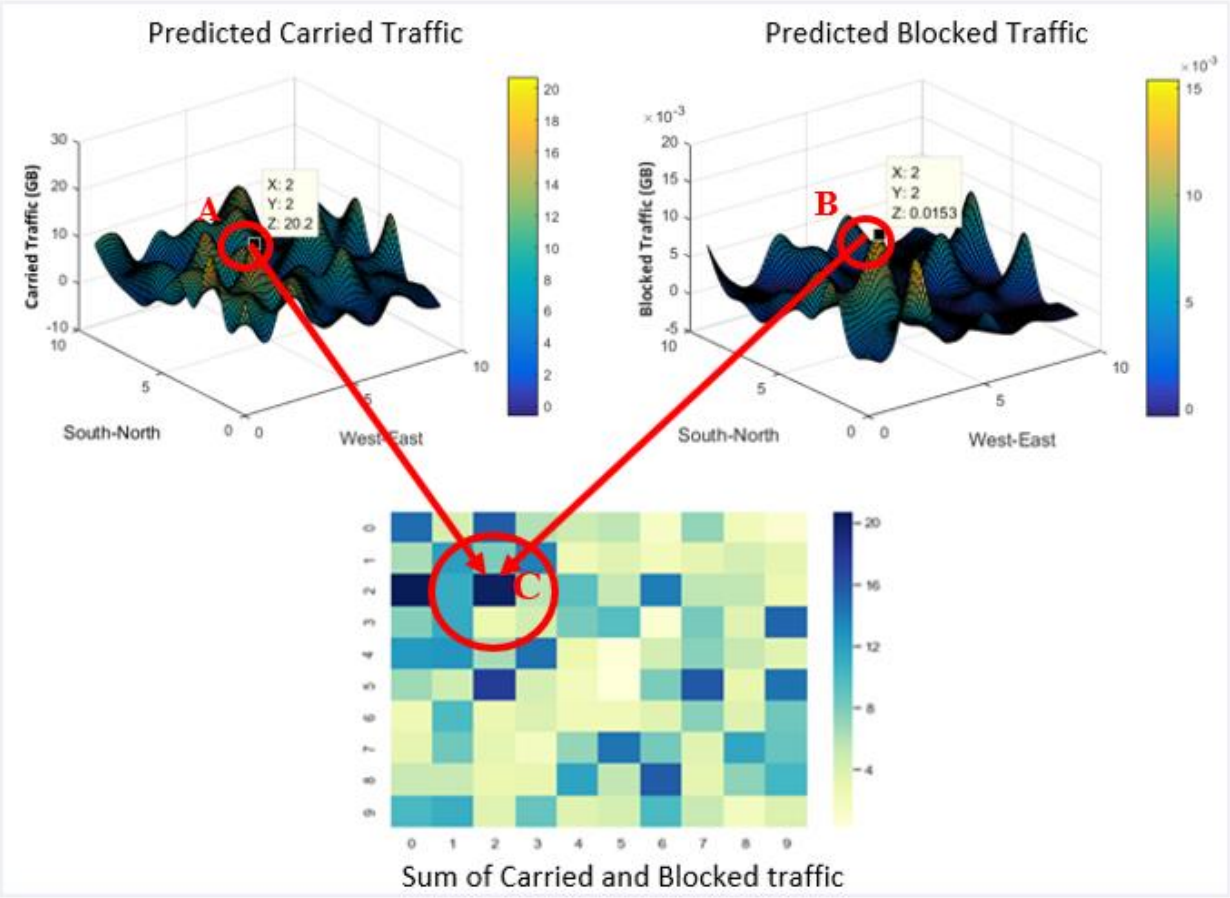


Figure 4-12 Carried Traffic VS Blocked traffic, on April 22

In this kind of situation, network optimization can be done using load balancing after doing a proper analysis of the network configuration. So, it helps to minimize the blocked data traffic on the specified grid for short term relief on the network and also gives insight on which area that needs urgent expansion in the long run. Furthermore, the information can be used as an input for minimizing congestion using dynamic price setting, and for energy-saving mechanism using dynamic cell zooming.

5. Conclusion and Recommendation

5.1. Conclusion

In this work, spatiotemporal mobile data traffic demand for the UMTS network is predicted using one of the deep neural network model ConvLSTM. The model capability, which captures the spatial complexity and temporal dynamics of the dataset, helps to improve prediction accuracy. After a twenty-four-hour observation of historical data, the model can predict the next six-hour with an RMSE value of 1.37. During the experiment, the model is tested using different hyperparameters and ConvLSTM model with three-layer, thirty-two filter, and 3*3 kernel performs better than the others, selecting the appropriate window size also improves the performance of the prediction model by delivering more information on convolution process.

Additionally, observing the predicted data traffic demand concerning blocked data traffic at a given time and space give significant insights for optimization, which facilitates the analysis process by detecting the blocked traffic area, and makes a better decision like load balancing.

Finally, this research delivers a significant insight into the future mobile data traffic demand at a given time and space, which can be used as an input for different purposes such as

- For the inhomogeneous mobile data traffic distribution, in which being aware of the growing data traffic demand at the area level is critical. It also helps to facilitate the planning process by localizing high traffic areas.
- The network optimization process needs to know the network status from different perspectives. One of them is getting the area level traffic growth and it is provided from this research. So, by implementing network optimization techniques based on space and time, the network performance will be optimized before the congestion occurs.
- For efficient management of network resources, like implementing energy-efficient resource utilization using dynamic cell zooming technique. To this end, knowing area level traffic load helps to control cell size accordingly.

- To facilitate the implementation of dynamic pricing mechanize; which is used for congestion control or encouraging the customer's usage behavior on the area where the data traffic demand is lower.

5.2. Recommendation

Investigating user-level data traffic usage behavior provides better information for mobile data traffic demand prediction at a specific time and space. So, it is recommended to get data traffic in small grid resolution and then implement spatiotemporal prediction with this resolution. It helps to identify hotspots and provide better predicted mobile data traffic.

References

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021," Cisco, 2017.

[2] Ethiotelecom, "Performance Report System," 2018.

[3] Guangshuo Chen, Sahar Hoteit, Aline Carneiro Viana, Marco Fiore and Carlos Sarratu, "The Spatiotemporal Interplay of Regularity and Randomness in Cellular Data Traffic," in LCN 2017 - The 42nd IEEE Conference on Local Computer Networks, Singapore, Oct 2017.

[4] Junbo Zhang, Yu Zheng, Dekang Qi, Ruiyuan Li and Xiuwen Yi, "DNN-Based Prediction Model for Spatial-Temporal Data," ACM ISBN, 2015.

[5] Samuel Medhn, Bethelhem Seifu, Amel Salem and Dereje Hailemariam, "Mobile Data Traffic Forecasting in UMTS Networks Based on SARIMA Model: The Case of Addis Ababa, Ethiopia," IEEE Africon 2017 Proceedings, 2017.

[6] Getinet Tesfaye, "Hybrid SARIMA-ELM- based Data Traffic Forecasting The Case of UMTS Network in Addis Ababa, Ethiopia," MSC. Thesis, AAIT, 2018.

[7] Yosef Abera and Dereje Hailemariam, "Spatio-temporal Mobile Data Traffic Modeling Using Fourier Transform Techniques," 2018 International Conference on Information and Communication Technology Convergence (ICTC), IEEE, 2018.

[8] Chih-Wei Huang, Chiu-Ti Chiang and Qiuhui Liy, "A Study of Deep Learning Networks on Mobile Traffic Forecasting," IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), 2017.

[9] Fengli Xu,Yuyun Lin, Jiaxin Huang, Di Wu, Hongzhi Shi, Jeungeun Song and Yong Li, "Big Data Driven Mobile Traffic Understanding and Forecasting: A Time Series Approach," IEEE Transactions on Services Computing, 2016.

[10] Guangshuo Chen, "Spatiotemporal Individual Mobile Data Traffic Prediction," HAL, February 2018 .

[11] Chaoyun Zhang and Paul Patras, "Long-Term Mobile Traffic Forecasting Using Deep Spatio-Temporal Neural Networks," Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing, 2018.

[12] Xu Wang, Zimu Zhouy, Zheng Yang, Yunhao Liu and Chunyi Peng, "Spatio-Temporal Analysis and Prediction of Cellular Traffic in Metropolis," IEEE Transactions on Mobile Computing, 2017.

[13] S.S. Riaz Ahamed, "comprehensive analysis of high performance Universal Mobile Telecommunication System (UMTS) in the wireless and mobile communication world," Journal of Theoretical & Applied Information Technology 4.11, 2008.

[14] Harri Holma and Antti Toskala , "HSDPA/HSUPA for UMTS: high speed radio access for mobile communications," John Wiley & Sons, 2007, pp. pp.1-9.

[15] Harri Holma and Antti Toskala, "WCDMA for UMTS," Third ed., Wiley & Sons, 2004, pp. 75-98.

[16] Xi Li, "Radio Access Network Dimensioning for 3G UMTS," Ph.D. dissertation, Hunan, China, 2009.

[17] Telecom-knowledge, "UMTS Radio Network KPI," Telecom-knowledge, [Online]. Available: https://telecom-knowledge.blogspot.com/2014/12/umts-radio-network-kpi.html. [Accessed 15 1 2020].

[18] Kreher Ralf, "UMTS performance measurement," A Practical Guide to KPIs for the UTRAN Environment, 1st ed. Chicester, UK: John Wiley & Sons, 2006.

[19] Latif Ullah Khan, "Performance Comparison of Prediction Techniques for 3G Cellular Traffic," IJCSNS International Journal of Computer Science and Network Security, vol. 7 No.2, February 2017.

[20] Manish R. Joshi and Theyazn Hassn Hadi, "A Review of Network Traffic Analysis and Prediction Techniques," arXiv preprint arXiv:1507.05722, 2015.

[21] Zhenglei Yi, Xin Dong, Xing Zhang and Wenbo Wang, "Spatial Traffic Prediction for Wireless Cellular System Based on Base Stations Social Network," in Annual IEEE Systems Conference (SysCon), 2016.

[22] Peter Jeffcock, "What's the Difference Between AI, Machine Learning, and Deep Learning?," Oracle, 11 July 2018. [Online]. Available: https://blogs.oracle.com/bigdata/difference-ai-machine-learning-deep-learning. [Accessed 8 10 2019].

[23] MathWorks, "What Is Deep Learning ?," MathWorks, [Online]. Available: https://www.mathworks.com/discovery/deep-learning.html. [Accessed 8 10 2019].

[24] Michael Nielsen, "Neural Networks and Deep Learning," San Francisco, CA, USA:: Determination press, vol. 25, 2015.

[25] vikashraj luhaniwal, "Forward propagation in neural networks — Simplified math and code version," towardsdatascience.com, 7 May 2019. [Online]. Available: https://towardsdatascience.com/forward-propagation-in-neural-networks-simplified-math-and-code-version-bbcfef6f9250. [Accessed 8 10 2019].

[26] Benjamin Sautermeister, "Deep Learning Approaches to Predict Future Frames in Videos," MSC. Thesis, TECHNISCHE UNIVERSITÄT MÜNCHEN, 2016.

[27] M. Mattheakis and P. Protopapas, "Recurrent Neural Networks: Exploding, Vanishing Gradients & Reservoir Computing," 2019.

[28] Chaoyun Zhang, Paul Patras, and Hamed Haddadi, "Deep learning in mobile and wireless networking: A survey," IEEE Communications Surveys & Tutorials, 2019.

[29] Keiron O'Shea and Ryan Nash, "An introduction to convolutional neural networks," arXiv preprint arXiv:1511.08458, 2015.

[30] He, Mingyi, Bo Li, and Huahui Chen, "Multi-scale 3d deep convolutional neural network for hyperspectral image classification," In 2017 IEEE International Conference on Image Processing (ICIP), pp. 3904-3908, IEEE, 2017.

[31] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in Proceedings of the IEEE international conference on computer vision, 2015.

[32] Herbert Jaeger, "Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the" echo state network" approach," vol. Vol. 5, Bonn: GMD-Forschungszentrum Informationstechnik, 2002.

[33] Christopher Olah, "Understanding LSTM Networks," colah's blog.Github, 27 August 2015. [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/. [Accessed 21 9 2019].

[34] Ayako Mikami, "Long Short-Term Memory Recurrent Neural Network Architectures for Generating Music and Japanese Lyrics," Ph.D., Computer Science Department, Boston College, 2016.

[35] Alex Graves, "Generating sequences with recurrent neural networks," arXiv preprint arXiv:1308.0850, 2013.

[36] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong and Wang-chun Woo, "Convolutional LSTM network: Amachine learning approach for precipitation nowcasting," Advances in Neural Information Processing Systems, 2015.

[37] François Chollet and others, "Keras," 2015.

[38] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv:1502.03167, 2015.

[39] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, and Ghemawat S, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," arXiv preprint arXiv:1603.04467, 2016.

[40] Tianfeng Chai and Roland R. Draxler., "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature," Geoscientific model development, 7.3, pp. 1247-1250, 30 June 2014.

[41] Colin D Lewis, "Industrial and business forecasting methods," London, Butterworth Scientific, 1982.

[42] Diederik P. Kingma and and Jimmy Lei Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.