Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

Telecommunication Engineering Graduate Program

Video Streaming Data Traffic Prediction by Using Long Short Term Memory (LSTM) Model: In the case of UMTS Network in Addis Ababa

By
Begameder Tamene

A Thesis Submitted to School of Electrical and Computer Engineering, in Partial Fulfillment of the Requirements for the Degree of Masters of Science in Telecommunications Engineering

Advisor: **Mesfin Kifle (PhD)**

February 2020

Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

Approval of the Thesis

Submitted By: - **Begameder Tamene**

Thesis Title: Video Streaming Data Traffic Prediction by Using Long Short Term Memory (LSTM) Model: In the case of UMTS Network in Addis Ababa.

The final reading approval of the thesis is granted by:

Mesfin Kilfe (PhD)	_____
Advisor	Signature
Ephraim Teshale (PhD)	_____
Examiner	Signature
Surafel Lemma (PhD)	_____
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.

Begameder Tamene _____

Name Signature

Place: Addis Ababa

Date of Submission: _____

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

Mesfin Kifle (PhD) _____

Advisor's Name Signature

Abstract

Predictive analysis of mobile network traffic is fundamental for the next-generation cellular network. Proactively knowing user demand allows telecom systems to perform optimal resource allocation. Nowadays, telecom companies face a network congestion problem; this problem results in longer delays, drastic jitter, and excessive packet losses. As a result, the quality of service (QoS) of networks deteriorates, and the quality of experience (QoE) perceived by end-users will be unsatisfied. As a solution, different researchers used statistical and neural network models for the prediction of video streaming data traffic. However, these models did not incorporate self-similarity and long term dependence characteristics of the video streaming data traffic. So, this study aims to predict the video streaming data traffic by using the Deep Learning, Long Short Term Memory (LSTM), model which incorporates self-similarity and long term dependence.We have reviewed various kinds of literature, conference papers, journals, white papers, and books related to the prediction of video streaming data traffic to achieve the objective of this study. Ten months of data (from October 2018 to July 2019) of video streaming data traffic information from five Radio Network Controllers (RNCs) of the Universal Mobile Telecommunication System (UMTS) network in the city of Addis Ababa (A.A) is collected. Finally, this research work result indicates that the LSTM model has 57.8% of MAE improvement of forecasting error compared to the hybrid model, i.e., Seasonal Auto-Regression Integrated Moving Average (SARIMA) and Extreme Learning Machine (ELM) model, which has the second lower error. The overall results of this research work demonstrate that the LSTM model is an effective method for predicting video streaming traffic to reflect temporal patterns. Such accuracy is vital to provide a better dynamic resource allocation for video streaming traffic.

Keywords: deep learning, forecasting, self-similarity, long term dependency, LSTM model, SARIMA model, ELM model

Acknowledgment

First and foremost, I give thanks to God for his absolute protection and ability to do this work. I am grateful to my company, ethio telecom, and Addis Ababa University's School of Electrical and Computer Engineering, Graduate Program in Telecom Engineering, for making it possible for me to study here.

I want to thank everyone who supported me throughout this thesis. Specifically, I would like to thank my advisor Dr. Mesfin Kifle for his invaluable support for advice and giving directions. A special thank you to my beloved family back home, whose love and support are my motivation and the reason I stand here today.

Finally, I would like to thank the Ethio telecom staff for providing me the data for the UMTS data traffic used as an input for the thesis.

Table of Contents

List of Figures

Acronyms

3G	Third Generation
A.A	Addis Ababa
ANNs	Artificial Neural Networks
ANFIS	Adaptive Network Fuzzy Inference System
AR	Auto Regressive
ARIMA	Autoregressive Integrated Moving Average
BPTT	Back Propagation Through Time
BS	Base Station
CAGR	Compound Annual Growth Rate
CN	Core Network
CS	Circuit Switch
ELM	Extreme Learning Machine
FFMLP	Feed Forward Multi-Layer Perception
GGSN	Gateway GPRS Support Node
GMSC	Gateway Mobile Switching Center
GSM	Global System for Mobile Communication
GPRS	General Packet Radio Service
GTP	Growth and Transformation Plan
HD	High Definition
HLR	Home Location Register
HTTP	Hyper Text Transfer Protocol
LTD	Long Term Dependence
LTE	Long Term Evolution
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MASE	Mean Absolute Square Error
ME	Mobile Equipment
ML	Machine Learning
MLP	Multi-Layer Perception
MPEG	Moving Picture Experts Group

MS	Mobile Station
MSC	Mobile Switching Center
MT	Mobile Terminal
MWM	Multifractal Wavelet Model
PS	Packet Switch
QoE	Quality-of-Experience
QoS	Quality of Service
RBFN	Radial Basis Function Networks
RMSE	Root Mean Square Error
RNC	Radio Network Controller
SARIMA	Seasonal Autoregressive Integrated Moving Average
SGSN	Serving GPRS Support Node
SNS	Social Networking
TAF	Terminal Adaptation Functions
TE	Terminal Equipment
UE	User Equipment
UMTS	Universal Mobile Telecommunication System
USIM	Universal Subscriber Identity Module
UTRAN	UMTS Terrestrial Radio Access Network
VLR	Visitor Location Register
WCDMA	Wideband Code Division Multiple Access

Chapter One
Introduction

This section aims to give some background information about the trend of the global and Addis Ababa mobile data traffic network. Moreover, the chapter discusses the statement of the problem and related work about the prediction of video streaming data traffic. The objective and method of the study also described.

1.1 Background

A major trend in mobile networks over the last few years is related to the exponential increase of mobile devices, such as smartphones and tablets, with multiple heterogeneous wireless interfaces[1]. Besides, the increase in the number of devices accessing the cellular network, emerging social networking platforms such as Facebook and Twitter have further added to the mobile data traffic[2]. Figure 1.1 shows mobile data traffic expected to grow to 77 Exabytes per month by 2022; it means a sevenfold increase over 2017. Mobile data traffic will grow at a Compound Annual Growth Rate (CAGR) of 46 percent from 2017 to 2022 [3]. According to Cisco's Visual Networking Index, mobile video traffic accounted for 59 percent of total mobile data traffic in 2017. This index estimated that videos would constitute more than three-fourths of the world's mobile data traffic by the year 2021.

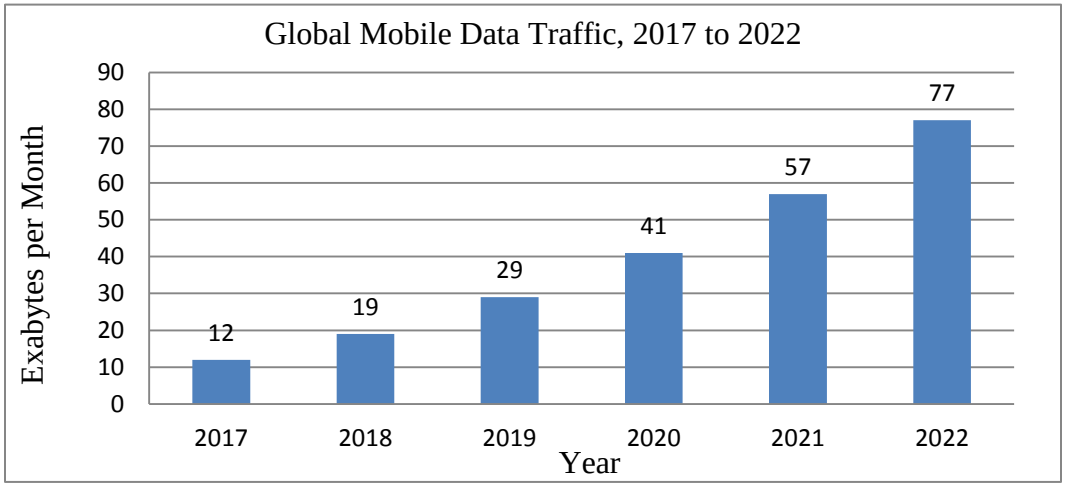

Figure 1.1: Global Mobile Data Traffic, 2017 to 2022 [3]

Such massive growth in video streaming traffic will lead to a tremendous amount of stress on the video delivery infrastructure. Therefore, the network service providers need to perform a careful

and optimal utilization of the available resources for video streaming while maintaining an acceptable level of QoE for video users.

Video streaming is a type of media streaming, in which the data from a video file is delivered continuously via the Internet to a remote user [4]. It allows a video to be viewed online, without being downloaded on a host computer or device. Video streaming works on data streaming principles, where all video file data is compressed and sent to a requesting device in small chunks. The size of each data stream depends on various factors, including actual file size, bandwidth speed, and network latency. In turn, the user or client player decompresses and displays the streamed data, allowing a user to begin viewing the file before the entire video file is received.

Because of smartphone penetration, telecom data services users increased. This truth is also working for the Ethio telecom customer. As shown in Figure 1.2 from the total data traffic, around 60% of the data comes from the smartphone.

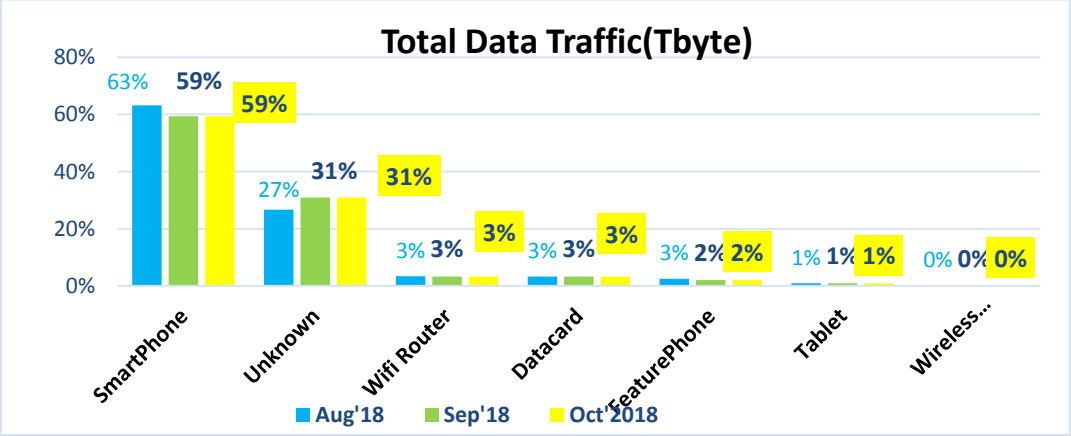

Figure 1.2: Total Data Traffic based on Device Usage from Aug 2018 - Oct 2018

In Ethio telecom from the entire data traffic that recorded in the three months (August 2018 to October 2018), the video streaming applications contribute to a significant share of Internet traffic. As shown in Figure 1.3, the three months (August 2018 to October 2018) of the total data traffic of the company categorized by level of application. Among these, the graph shows 24 % and 22% share go-to streaming and Social Networking (SNS) traffic, respectively.

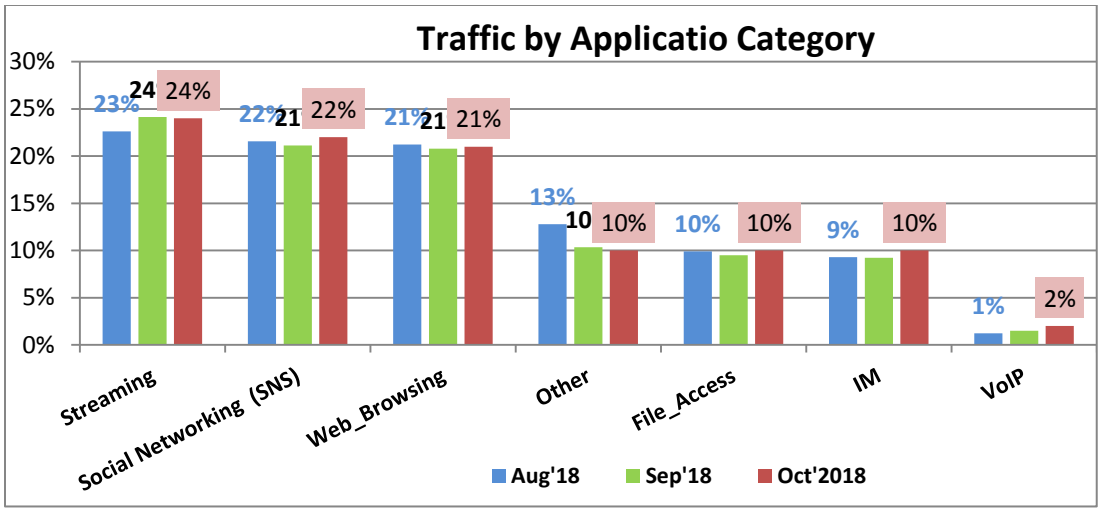

Figure 1.3: Total Data Traffic based on Application from Aug 2018 - Oct 2018

Video is going to be the dominant Internet traffic, and the most popular standard used to transport and view video is Moving Picture Experts Group (MPEG)[5]. The MPEG traffic is Variable Bit Rate (VBR) traffic & in the form of a time-series representing frame sizes. This time-series is extremely noisy, and analysis shows that it has a very long-range time dependency, making it even harder to predict than any typical time-series. Traffic prediction is essential in enhancing reliable operation over these networks.

Extracting previous traffic demand information from big data to manipulate and use it to forecast the future demand is becoming a trend with telecom service providers globally. The efficient prediction of traffic generated by multimedia sources is an essential part of traffic and congestion control procedures at the network edges. The capability to predict future video streaming data traffic will enable a more effective bandwidth allocation mechanism.

Telecommunications industries are growing very fast, and the ability to determine future trends is an important task. There is a problem in bridging the gap between the known data and probable value in the future due to the lack of reliable input data for forecasting and adequate model. There are different statistical forecasting models such as Moving Average (MA), exponential or linear regression Auto-Regression Moving Average (ARMA), Auto-Regression Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), Fractional Auto Regression Integrated Moving Average (FARIMA), Support Vector Machine (SVM), Multilayer Perceptron (MLP) and Neural Network (NN). These models can be used separately or in

combination for reasonable prediction, but these models do not incorporate all the property of the data traffic. As a result, the time series of data traffic not forecasted accurately. So, it is necessary to find new ways to minimize the effects of poor forecasts [6]. Ethio Telecom, the sole telecom service provider in Ethiopia, provides national and international telecommunications services, using communication media of satellite, optical fiber, microwave, multiple access radius, tiny aperture, ultra-high frequency, and very high frequency. Currently, it provides fixed-line telephony, mobile telephone, and Internet and multimedia services [7].

The 3G wireless service had been launched in 2013 in A.A and then extended to the other major cities of Ethiopia. 3G services are becoming more popular, and the traffic over the network is overgrowing. With its massive expansion plans, Ethio telecom had 58.08 million subscribers in a year of 2016/17. According to the Ethiopian Government's second Growth and Transformation Plan (GTP2), this number expected to rise to 103.66 million by the year 2020[8]. This massive expansion plan will require efficient network planning, deployment, and operations [9].

1.2 Statement of the Problem

With a rapid popularization of the Internet, the network scale, users, and internet applications have drastically increased. In Ethio telecom video streaming applications contribute a significant share of Internet traffic. The company's ten months (October 2018 to July 2019) report indicates that higher percentage share of total data traffic goes to video streaming.

On the Internet, network congestion is becoming an intractable problem. Congestion results in longer delay, drastic jitter, and excessive packet losses. As a result, the QoS of networks deteriorates, and then the QoE perceived by end-users will not be satisfied.

To overcome these problems, using past observations to predict future network traffic is a crucial step to understand and plan the capacity of the Ethio telecom infrastructure. Most researchers use statistical and neural network forecasting models to predict video streaming data traffic. An ideal data traffic prediction model needs to have the ability to capture prominent traffic characteristics, such as long-term dependence and self-similarity in a larger time scale, multifractal on a smaller time scale [10]. However, these statistical and neural network models do not incorporate all the characteristics of video streaming, i.e., long term dependency, and self-similarity.

1.3 Objective

1.3.1 General Objective

The main objective of the study is to predict UMTS video streaming data traffic by using the Deep Learning, LSTM, model in Addis Ababa city.

1.3.2 Specific Objectives

The following specific objectives will help to achieve the general aim of the study

- To review appropriate forecast models for video streaming data traffic.
- Collecting, pre-process, analyzing, and visualize video streaming data traffic of Ethio telecom
- Select specific Deep learning model that fits into the data
- Analyze the performance of the implemented model with forecast error metrics (RMSE, MAE, MAPE, and MASE).
- Select a model based on minimum forecast error
- Evaluate the selected deep learning model.

1.4 Methodology

To fulfill the general and specific objectives of this research, we follow the following methods:

Literature review

To achieve the objective of this study, various literature, conference papers, journals, white papers, and books related to the prediction of video streaming data traffic were reviewed. These help to understand the area and gain knowledge on how others have seen the problem and gone through it.

Data Collection

To provide measurement data based forecasting, ten months of video streaming data traffic daily, i.e., from October 2018 – July 2019, is extracted from the Huawei Smart Care Solution. The collected sample data, 304 days of data, pre-process, explore, and visualize the series. Then, the collected data categorized into two groups for the training and test purpose.

Tools and algorithms

We used different forecasting models, such as hybrid (SARIMA-MLP and SARIMA-ELM), and deep learning (LSTM) models. For evaluating the performance of each model; RMSE, MAE, MAPE and MSE metric are used. R statistical tools and Jupiter notebook used for the modeling and processing of the data. Additional tools like Microsoft Excel and SPPS software also used for the analysis purpose. Figure 1.4 demonstrated a summary of the methodology used.

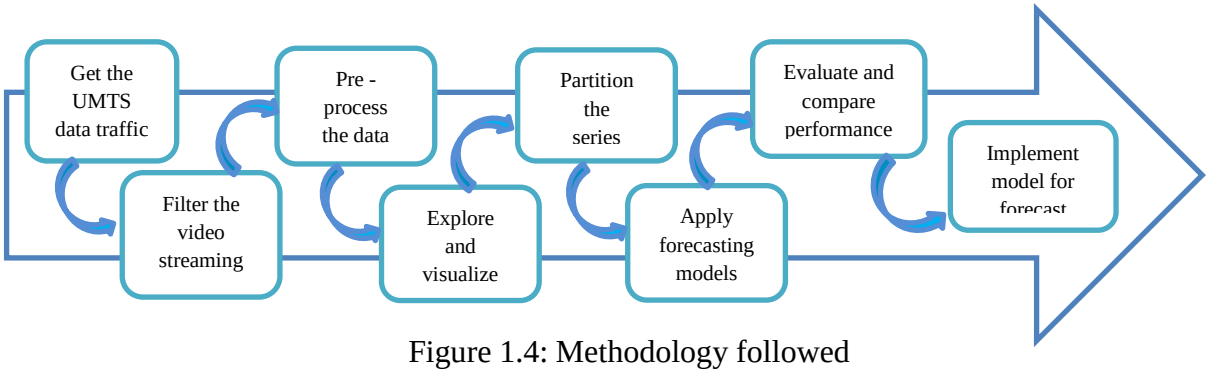

Figure 1.4: Methodology followed

1.5 Related work

There are many studies conducted on the prediction of video streaming data traffic. Some of them represented as follows as.

In the earliest work, network traffic prediction models use linear time series models, e.g., Auto-Regressive (AR) and ARIMA. The exponential decay of the autocorrelation function of these models gives them the ability to capture the short term dependence (STD) characteristics only. However, the traffic data exhibited a high degree of long term dependence (LTD) characteristics, in addition to STD. Thus, such models cannot characterize the network traffic well [11].

Another related model, known as FARIMA model, can capture both STD and LTD and has been used to model and predict traffic data. However, this model cannot capture the multifractal, which has been found in the network traffic on a small time scale. For this reason, another Multifractal Wavelet Model (MWM) has been introduced to solve this problem. MWM model can capture multifractal but cannot predict traffic [12].

In [13], the aim of the authors was to predicted the video stream for efficient bandwidth allocation of the video signal. They consider video stream prediction for application services like

video-on-demand, videoconferencing, video broadcasting, etc. The efficient prediction of traffic generated by multimedia sources is an essential part of traffic and congestion control procedures at the network edges. As a tool for the forecast, it used neural networks such as MLP, Radial Basis Function Networks (RBF), and Back Propagation Through Time (BPTT) neural networks. They also propose moving averages for the video time series processing. Finally, present the simulation results for each type of neural network together with final comparisons of each model, and they construct suitable neural networks for variable bit rate video prediction and evaluate results.

In the study[14], the authors have predicted the future frame of the video stream. It describer the single frame prediction problem, i.e., the information of previous frame sizes was used for the prediction of the next frame size of the sequence. As a tool, it used a Feed-Forward MLP(FFMLP) neural network. The author found self-similarity and non-linearity characteristics of internet traffic. The discovery of self-similarity and non-linearity of network traffic has brought challenges to traffic prediction. So, the author used the feed-forward neural network for performing nonlinear traffic prediction. Finally, the authors investigate the capability of the neural network for video traffic prediction, included the non-linearity characters of the traffic, but unable to add the features of self-similarity.

The study held on Thailand [15], the authors model, and predict high-definition (HD) video traces encoded with an H.264/AVC encoding standard. The results that the authors got based on the compilation of over 50 HD video traces. By using a simplified SARIMA model provides an accurate representation of HD videos and significant improvements in prediction accuracy. Such accuracy is vital to give a better dynamic resource allocation for video traffic. In addition, the authors provide a statistical analysis of HD videos, including both factor and cluster analysis, to support a better understanding of video stream workload characteristics and their impact on network traffic.

In the study presented [16], a neuro-fuzzy approach was used to predict the video traffic for MPEG-4-coded videos. The predictor was based on the Adaptive Network Fuzzy Inference System (ANFIS) to perform single-step predictions for the I(intra-frame), P(predictive), and B (bidirectional) frames. Short-term forecasts are also examined using smoothed signals of the video sequences. The authors evaluated the ANFIS prediction results by using long

entertainment and broadcast video sequences and compared them to those obtained using a linear predictor. Finally, it concludes that ANFIS is capable of providing accurate predictions than linear predictor and had the added advantage of being simple to design and to implement.

In the study [17], the authors classifying video content based on feature extraction using the cluster analysis into three specific content types of ‘slow movement,’ ‘gentle walking,’ and ‘rapid movement’ based on the combination of temporal and spatial feature extraction. Also, the authors investigate the combined effects of network and application parameters on end-to-end perceived video quality over wireless networks for three distinct content types. The authors develop two models for video quality estimation as (a) a hybrid video quality prediction model based on an ANFIS. (b) A regression-based model for the three content types. Finally, the authors used ANFIS to train the three neural networks for three distinct content types to predict the video quality based on a set of objective parameters. The authors used a testbed based on simulated network scenarios using a network simulator NS-2 with an integrated tool Evalvid. The authors observed that network-level parameters (link bandwidth and packet error rate) have a much bigger impact on video quality as expected compared to application-level parameters (frame rate and video send bitrate). The authors put the experiment result as follows. The correlation coefficient and RMSE for MOS scores were generally better than the decodable frame rate except in ‘gentle walking’ where Q results were better. The results confirm that the proposed ANFIS-based Artificial Neural Network (ANN) learning model is a suitable tool for video quality estimation for the most significant video streaming content types. The regression-based model gave a better prediction performance compared to that of ANFIS. Both the models were validated with video clips within the same content type with good prediction accuracy.

According to [18] SARIMA model was used for modeling and prediction of 4K video traces, encoded with an H.265 encoding standard. The analysis was used based on publicly available HD video traces in the R programming environment. One of the reasons that the author for choosing SARIMA models was because SARIMA models consider both non-seasonal and seasonal parts of data traces. The author used RMSE and MAE as the prediction accuracy. As a conclusion, the author found a high frame size variance was not taken into account on purpose, and this hurts both, 4K video traffic modeling and prediction. Therefore, the author confirms that

a frame size variance could have a negative impact on 4K video traffic modeling and prediction, especially if wide intervals are used in the learning phase.

According to [19], the main aim of the author is to model and forecast UMTS data traffic demand based on historical data collected from the live network, with a linear Univariate time series method, by using a hybrid SARIMA-ELM model to increase the accuracy of the forecast. For the linearity of the data, the author uses SARIMA models, and the errors of this SARIMA model fit has been extracted as residuals and modeled with ARNN, MLPNN, and ELMNN nonlinear models. Then the author used the hybrid of the linear and nonlinear model. Finally, the author used RMSE, MAE, MAPE, and MASE error metrics comparison for the hybrid SARIMA-ARNN, SARIMA-MLP, and SARIMA-ELMNN models. The SARIMA-ELMNN forecast model scores a minimum error value of these metrics and an average of 24.8 % error improvement than the SARIMA–MLP model, which has the next minimum value of the error metrics. For data analysis, the author used R statistical tools.

All reviewed papers dealt with forecasting of a video streaming data traffic, based on historical data collected. Most of the papers used a single linear or non-linear models; however combining models in a hybrid model increases a forecast performance as shown on [18]. But all the paper that used for forecasting video data traffic predication model do not incorporated the self similarity and LTD characteristics.

1.6 Scope and limitation

1.6.1 Scope of the Thesis

The scope of this study is to forecast the sampled mobile UMTS network video streaming data traffic using the best-fitted(LSTM) model. Besides this, compare the result with two different hybrid-forecasting models, SARIMA-MPL and SARIMA-ELM, and select the best one with the minimum prediction error. And also, this study addresses the importance of forecasting in resource allocation and network infrastructure capacity terms.

1.6.2 Limitation of the Thesis

There are lots of factors that can be considered and make the forecasting accuracy high, like customer behavior modeling and treating the data as multivariate. Because of the complexity and

shortage of time, the proposed model used only univariate data, i.e., the total volume of the video streaming data traffic.

1.7 Contribution of the Study

The Contribution of the study is by using deep learning (LSTM), with minimum forecast error, model to forecast A.A UMTS video streaming data traffic based on historical data collected. The proposed model increases the performance of forecasting in video streaming data traffic than machine learning and NN models because it incorporates additional features, i.e., self- similarity and long term dependency.

Thus, Ethio telecom will be benefited by using the LSTM model that captures both long-term dependence and self-similarity of the UMTS video streaming data traffic characteristics.

1.8 Thesis organization

The remaining of this thesis organized as follows: Chapter Two discusses the UMTS network and infrastructures. In Chapter Three, it describer the video streaming traffic architecture, different types of video streaming based on the existence on the internet, and the hindrance of the video streaming quality briefly. Chapter Four deals with the time series data traffic and different types of forecasting models. Chapter Five describes an experiment analysis, results, and discussion. Chapter Six, the last chapter, is all about conclusion, recommendations, and future works.

Chapter Two
Universal Mobile Telecommunication System Network

This chapter introduces the basic fundamental principles of UMTS networks, an overview of the UMTS architecture and its components briefly. The UMTS network architecture divided into three module: User Equipment (UE) - which is the final interface with the user, UMTS Terrestrial Radio Access Network (UTRAN) – which facilitate effective handover between Node Bs under the control of different RNCs, and Core Network (CN) - which is responsible for switching and routing calls and data connections to external networks. Finally, discuss the technology that Ethio telecom to use.

2.1 UMTS Network

The mobile telecommunication system introduced as a commercial cellular system for the public in the late 20th century[20]. Mobile networks have evolved through five generations. The first generation (1G) mobile wireless communication network was analog used for voice calls only. The second-generation (2G) is a digital technology and supports text messaging. The third-generation (3G) mobile technology provided a higher data transmission rate, increased capacity, and provide multimedia support. The fourth-generation (4G) integrates 3G with fixed internet to support wireless mobile internet, which is an evolution to mobile technology, and it overcomes the limitations of 3G. It also increases the bandwidth and reduces the cost of resources. 5G stands for 5th Generation Mobile technology and is going to be a new revolution in the mobile market, which has changed the means to use cell phones within very high bandwidth [21]. The number of mobile subscribers that growth over the last years has contributed to the evolution of mobile networks. From the early 1G analog mobile to the last implemented 5G system, it has been changed both in terms of services provided and technological complexity. In the latest generation of networks, communication is ubiquitous, and they provide users with a new set of services.

The UMTS is a 3G mobile communications system that provides a range of broadband services to the world of wireless and mobile communication users. It is a 3G mobile cellular system for networks that evolved from the 2G Global System for Mobile Communication (GSM) standard and is a component of the IMT-2000 standard set. UMTS offers higher spectral efficiency and

available bandwidth to mobile network operators, and this attributed to its use of Wideband Code Division Multiple Access (WCDMA) radio access technology[20]. UMTS efforts were initiated in the 1992 meeting of the International Telecommunication Union (ITU). Its original goal was to design a single 3rd generation air interface. Commercial networks were first deployed in early 2001 in Japan and later rolled out in Europe in 2002 [22].

The upgraded network form GSM to UMTS reuse several network elements, including the Home Location Register, Visitor Location Register, Mobile Switching Center, and the Authentication Center, to name some. However, a new Base Station Controller and Base Transceiver Station are required. In most cases, 3G and 2G networks will be made to operate side by side.

One of the critical functionalities of UMTS is the ability to provide services anywhere and anytime. In UMTS, mobile equipment can use for communication, entertainment, business, and all kinds of services. With an increase in the number of smartphones which has different applications setup, the data traffic demand is likely to be higher. This increment pushes researchers and network operators to develop a new plan for the existing infrastructure of the network.

With the wireless industry now moving from 3G to 4G, UMTS serves as the basis of the 3GPP's new set of radio technologies, known as Long Term Evolution (LTE).

2.2 UMTS Network Architecture

The UMTS network architecture divided into three modules [23]: UE, UTRAN, and CN. The Radio Interface, Uu, connects the UE to the UTRAN; and the CN-UTRAN interface, Iu, connects the UTRAN to the CN. Figure 2.1 shows the UMTS architecture.

2.2.1 User Equipment

The UE is a major element of the overall 3G UMTS network architecture. It forms the final interface with the user. The UE aggregates the Mobile Equipment (ME) and the Universal Subscriber Identity Module (USIM). The ME is the single or multimode Mobile Terminal (MT) used for radio communication over the Uu interface. The USIM is a smart card that holds the

subscriber identity, performs authentication algorithms, stores authentication and encryption keys, and information needed at the terminal.

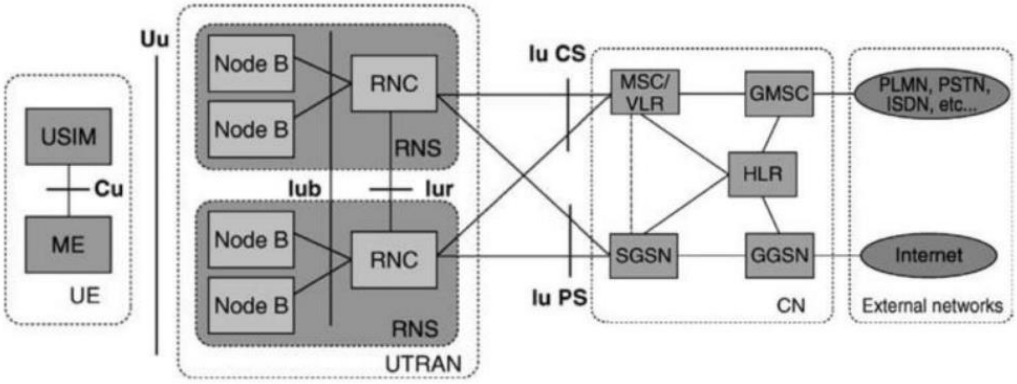

Figure 2.1: UMTS System Architecture[24]

Figure 2.2 gives a clear view of the UE component. The Cu interface enables the communication between the USIM and the ME. Mobile equipment has three parts such as

- Terminal Adaptation Functions (TAF)
  - It is service dependent and flows control/rate adaptations
  - It is the mapping of terminal requests on network capabilities
- Terminal Equipment (TE)
  - Its support for end-to-end application functions necessary for the operation of the access protocols by the user, e.g., a laptop
- Mobile Termination (MT)
  - It is a function for radio transmission and management of the radio interface, e.g., the handset.

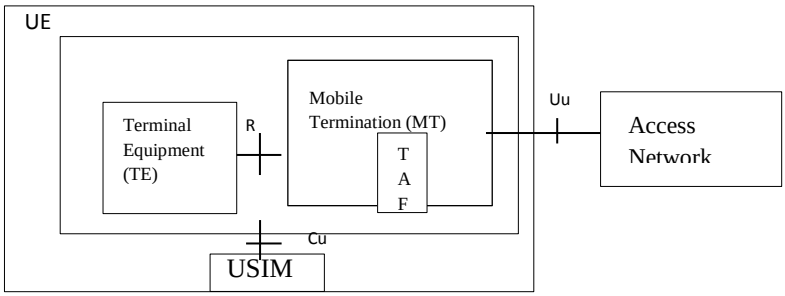

Figure 2.2: User Equipment

2.2.2 UMTS Terrestrial Radio Access Network

The UTRAN is composed of several Radio Network Subsystems (RNS). Each RNS includes a RNC and Node B's. The Iub interface connects the RNC to the Node B's. Node B, which represents the Base Station (BS), converts the data flow between the Iub and the Uu interfaces and participates in the Radio Resource Management (RRM). It performs air interface processing (channel coding, rate adaptation, spreading, synchronization, and power control). The RNC controls the Node B's connected to it and also executes the RRM. The Iur interface enables the connection between RNCs. The RRM assures the outer loop power control, packet scheduling, and handover control. The UTRAN functions are handover, provision of radio coverage, RRM, and control, system access control, security, and privacy. Figure 2.3 below shows the structure of UTRAN.

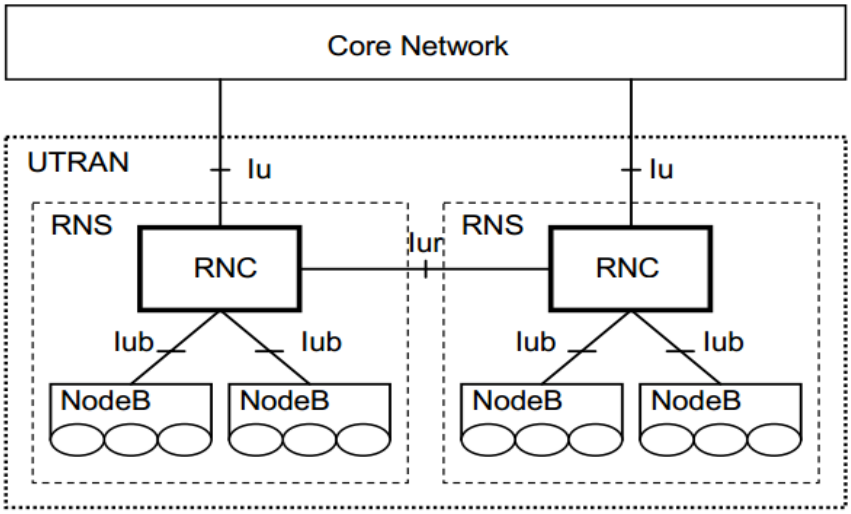

Figure 2.3: UTRAN Architecture [25]

2.2.3 Core Network

The CN aggregates the Packet Switch (PS) network and the Circuit Switch (CS) network. PS is responsible for switching and routing calls and data to external systems, and CS is responsible for the public switched telephone network. Some of the circuit-switched elements are Mobile Switching Center (MSC), Visitor Location Register (VLR), and Gateway MSC (GMSC). On the other hand, packet-switched parts are the Serving GPRS Support Node (SGSN) and Gateway

GPRS Support Node (GGSN). Both domains share some network elements like Home Location Register (HLR) and VLR. Below is a brief description of the CN elements.

- The HLR is a database where the operator subscriber's information is stored, such as allowed services, user location for routing calls, and preferences.
- The MSC/VLR is the switch (MSC) and database (VLR), which serves the UE in its location CS services.
- The GMSC is the gateway node between the circuit domain of the mobile network and the external network.
- SGSN is responsible for delivering packets to/from the MSs within its service area and communications with the GGSN. It also keeps track of the mobiles within its service area.
- GGSN acts as a logical interface to external packet data networks. It also maintains routing information used to tunnel the Protocol Data Unit (PDU) to the SSGN that is currently serving the Mobile Station.

The CN functions are mobility management, operations, administration, and maintenance; switching allowance; service availability; the transmission of MT traffic between UTRAN(s) and/or fixed network(s)[26].

Ethio telecom used the High-Speed Packet Access (HSPA) and evolved HSPA (HSPA+) technology of the UMTS network, which are available technologies in Addis Ababa. According to [27], HSPA refers to a set of technologies that enable operators to upgrade their networks to run at the speed of broadband networks. HSPA is a combination of two mobile protocols, High-Speed Downlink Packet Access (HSDPA) and High-Speed Uplink Packet Access (HSUPA) that extends and improves the performance of existing 3G mobile telecommunication networks using the WCDMA protocols. HSPA networks can support a maximum of 14.4 Mbps downlink speed and 5.7 Mbps of peak uplink speed of throughput per cell. HSPA+ supports data rates of up to 42 Mbps of throughput per cell. By using dual cell deployment and multiple inputs, multiple output architecture, HSPA+ networks can achieve maximum throughput of 168 Mbps overall for downloads and up to 22 Mbps for uploads.

2.3 Chapter Summary

In this chapter, to introduce some fundamental principles of UMTS networks. At first, to describe the evolution of the mobile telecommunication system followed by the introduction of UMTS system architecture and its components. UMTS networks have been deployed worldwide as of 3rd generation mobile communications systems. So UMTS offers higher spectral efficiency and available bandwidth to mobile network operators. UMTS provides a clear evolutionary path to high-speed packet access (HSPA). HSPA refers to the combination of high-speed downlink packet access (HSDPA) and high-speed uplink packet access (HSUPA).

Chapter Three

Video Streaming

This chapter describes the architecture of video streaming and its component briefly. It defines what video streaming means. It discusses how video streaming services can be offered over the Internet and also illustrates the fundamental problems that affect video streaming for the quality of service. Finally, it describes the services quality monitoring and analysis in Ethio telecom with the critical quality indicator (parameters).

3.1 Definition and Overview of Streaming Video

Video streaming is a type of media streaming, in which the data from a video file continuously delivered via the Internet to a remote user[4]. It is content sent in compressed form over the Internet and displayed by the viewer in real-time. When it uses a streaming video or streaming media, a Web user does not have to wait to download a file to play it. Instead, the media is sent in a continuous stream of data and play as it arrives. The user needs a player, which is a unique program that uncompressed and sends video data to the display and audio data to speakers. A player either can be an integral part of a browser or downloaded from the software maker Web site.

Major streaming video and streaming media technologies include Real System G2 from Real Network, Microsoft Windows Media Technologies (including its NetShow Services and Theater Server), and VDO. Microsoft approach uses the standard MPEG compression algorithm for video. The other approaches use proprietary algorithms. Microsoft technology offers streaming audio at up to 96 Kbps and streaming video at up to 8 Mbps (for the NetShow Theater Server). However, for most Web users, the streaming video will be limited to the data rates of the connection (for example, up to 128 Kbps with an ISDN connection). Microsoft streaming media files are in it's Advanced Streaming Format [28].

Video streaming is a best-effort delivery network. It means the network system does not provide any guarantee that data is delivered or that delivery meets any quality of service. Delivering audio or video content over the Internet can be achieved by two methods: progressive download and real-time streaming. If the content size is short, it used the progressive download method. In

this method, media content directly downloaded from a server into storage units of a client. However, in real-time streaming, client software plays media content without storing the content into any storage units. The client software is responsible for performing the media as it is delivered[29].

3.2 Video Streaming Architecture

Streaming video is a sequence of moving images that are sent in compressed form over the internet and display by the viewer as they arrive. Until recently, video on the web was primarily a download-and-play technology. Now a day, streaming video files begin playing almost immediately, while data is being sent, without having to wait for the whole file to download.

As shown in Figure 3.1, the architecture of video streaming divided into six areas, such as media compression, application-layer QoS control, media distribution services, streaming servers, media synchronization at the receiver side, and streaming media protocols[30].

Media compression: Video compression reduces the irrelevancy in the video signal by only coding video features that are perceptually important. It follows a standard for multimedia content that encodes the content with a specific play rate.

Application-layer QoS control: It involves congestion control and error control, which are implemented at the application layer.

Media distribution services: After the adaptation by application-layer QoS control module, the transport protocols packetize the compressed bit-streams and send the video/audio packets to the Internet.

Streaming servers: It plays a vital role in providing streaming services. Streaming servers are required to process multimedia data in real-time, support VCR like functions, and synchronously retrieve media components.

Protocols for streaming media: Streaming protocols provide means to the client and the server for services negotiation, data transmission, and network address. According to the functionalities, the protocols directly related to Internet streaming video can be classified as network-layer protocol, transport protocol, and session control protocol.

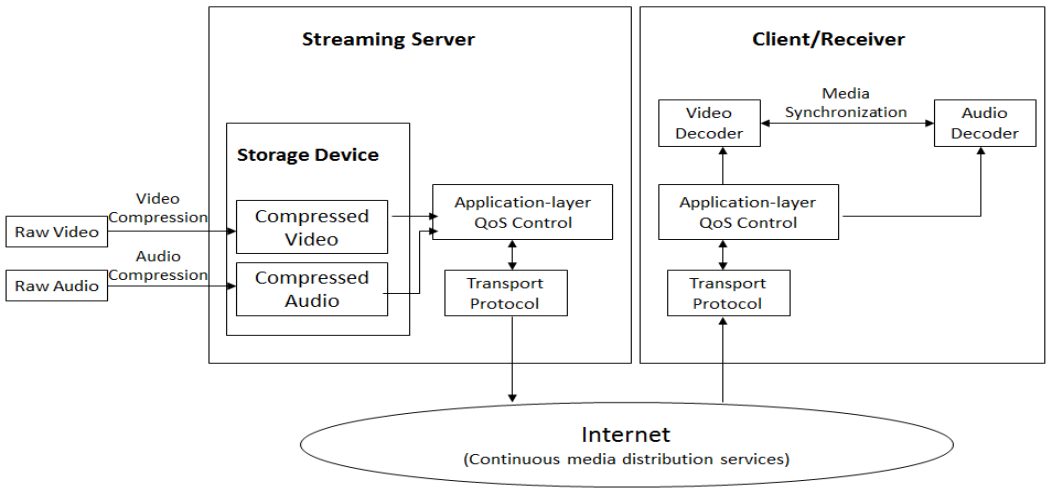

Figure 3.1: An Architecture for Video Streaming [31]

Media synchronization at the receiver side: With media synchronization mechanisms, the application on the receiver side can present various media streams in the same way as they captured initially.

3.3 Existing Streaming Networks

There are three essential means in which a streaming service can offer over the Internet, such as web-based distribution, On-demand Multimedia Streaming, and live video streaming.

3.3.1 Web-Based Distribution

It is the most frequently used technique to serve small streaming content. When a large number of user requests arrive, a Web-based content distribution architecture suffers from server overloading. Hence, appropriate schemes are required to manage the server loads effectively. Content caching and replication techniques direct the workload away from possibly overloaded origin Web servers to deal with Web performance and scalability from the client-side and the server-side, respectively[32].

3.3.2 On-Demand Multimedia Streaming

It enables immediate distribution of video streams to users regardless of the time at which the service request arrives with other ongoing streaming sessions. Typically, these video files are stored in a set of central video servers, and distributed through high-speed communication

networks to geographically-dispersed clients. Upon receiving a client‘s service request, a server delivers the video to the client as an isochronous video stream.

3.3.3 Live Video Streaming

It enables users to watch several TV channels through the Internet simultaneously. In live streaming, video streams generate at the same time as it is being downloaded and viewed by the clients. The popular live video streaming service is Internet Protocol Television (IPTV). With the extensive acceptance of residential broadband access and the progress of video compression technologies, IPTV may be the next popular Internet application.[33].

3.4 Basic Problems in Video Streaming Services Quality

There are some fundamental problems such as Bandwidth, Delay jitter, or Loss rate, that affect video streaming services quality. Because video streaming services over the Internet use best-effort delivery, there are no guarantees on bandwidth, delay jitter, or loss rate. Therefore, a key goal of video streaming services quality is to design a system to deliver high-quality video over the Internet reliably[34].

3.4.1 Bandwidth

The bandwidth available between two points on the Internet is generally unknown and time-varying. If the sender transmits faster than the available bandwidth, then congestion occurs, packets are lost, and there is a severe drop in video quality. If the sender transmits slower than the available bandwidth, then the receiver produces sub-optimal video quality. The goal to overcome the bandwidth problem is to estimate the available bandwidth and then match the transmitted video bit rate to the available bandwidth.

3.4.2 Delay Jitter

The end-to-end delay that a packet experiences may fluctuate from packet to packet. This variation in end-to-end delay referred to as the delay jitter. Any late frames resulting from the delay jitter can produce problems in the reconstructed video, e.g., jerks in the video. This problem typically addressed by including a playout buffer at the receiver. While the playout buffer can compensate for the delay jitter, it also introduces additional delay.

3.4.3 Loss

Different types of losses may occur, depending on the particular network under consideration. For example, packet loss afflicted wired packet networks such as the Internet, where an entire packet is lost. On the other hand, wireless channels afflicted by bit errors or burst errors. It can have a very destructive effect on the reconstructed video quality. To overcome this impact, to use designed with error control. We can use the following four approaches for error control, such as forward error correction (FEC), retransmissions, error concealment, and error-resilient video coding[35].

3.5 Metrics To Measure Video Streaming Performance

There are five most important metrics to measure the performance of video streaming. There are:

3.5.1 Bit Rate

Bit rate is one of the most important metrics to measure the performance of video streaming. The bit rate indicates the number of bits of a video that transmitted over a specified period. Bit rate has a direct impact on the play rate. The videos which are appealing to a larger audience will have a higher play rate when compared to the other. A higher average bit rate means a higher quality image (for a given screen resolution).

3.5.2 Buffer Fill

Buffering helps to understand the duration of time a user waits before a specific video starts to play. If they are staring at that rotating circle for too long, they may leave before even a frame flashes on their retinas.

3.5.3 Lag Length

This metric helps to understand the user's experience when they begin watching a video. The lag length is a time to spend while the video buffers until it begins to play. The lag length should not be much longer than the buffer fill time, but if it does happen, that means the video streaming quality is extremely bad and needs a performance check.

3.5.4 Play Length

It is the total amount of data consumed by the viewer, which includes every second, minute, and hour of the video streaming. This metric is equally essential to understand for capacity and infrastructure planning, and also helps you estimate peak data volumes and the overall demand for streamed data.

3.5.5 Lag Ratio

Waiting time is technically called lag ratio, which measured by calculating the overall waiting time in the video over the time spent watching the video. The lag ratio is never zero because the initial buffer fill that comes before the video plays remains, even if there is never a single buffer through the video.

These five metrics help improve the quality of the video so that the conversion rate can be boosted in the long run. Consumers want the kind of video streaming service, which is consistent, high quality, and easy to access, with less buffering. A geographically scattered viewer will have a direct impact on the performance of the video streaming, which is why an up-to-date load testing platform required for streaming video. The infrastructure of the video streaming needs to be designed so consumers get to watch the videos they want, on whatever device they want, and at the highest quality, they expect.

3.6 Video Streaming Quality Monitoring and Analysis in Ethio telecom

Ethio telecom uses Huawei's PS probe system (NetProbe3010), Service & Experience Quality (SEQ) Analyst V200R002C00, and Visual IP system for streaming quality monitoring and analysis purpose. SEQ Analyst is dedicated to efficiently managing service quality and network performance, rapidly process customer complaints, and support experience marketing. The Huawei Visual IP system used for IP bearer network monitoring and demarcation.

For Streaming quality monitoring, the SEQ Analyst displays the locations where service issues occur on a GIS map. It updates the Streaming Key Quality Indicators (KQI) data every 5 minutes for specified locations and generates an alarm when a predefined alarm threshold is reached. It enables near-real-time monitoring of Streaming KQIs.

For Streaming quality analysis, the SEQ Analyst can analyze Streaming KQIs from multiple dimensions (including time, location, and access technology), identify major service issues, and troubleshoot them in a timely manner. With such data, you can take effective measures to improve Streaming service quality and user experience.

Ethio telecom uses five KQI parameters for the video streaming service quality. They are:-

Video Streaming Start Success Rate: This KQI indicates the rate at which multimedia files are successfully played on a web page after a user clicks the play button.

Video Streaming Start Delay: This KQI indicates the amount of time a user waits until a multimedia file is played after clicking the play button on a web page.

Video Streaming Stall Frequency: This KQI indicates the number of times a streaming media file stalls within one minute when the file is being played.

Video Streaming Stalled Time Rate: This KQI indicates the percentage of the total stall time out of the total length of the played multimedia file (audio or video).

Video Streaming Plays Disconnected Rate: This KQI indicates the rate at which the downloading of a multimedia file stops or restarts before the multimedia buffer is fully occupied because the client or server fails to process the multimedia service efficiently or the network transmission quality is poor.

3.7 Chapter Summary

This chapter started by defining what video streaming means based on reviewing literature. It described the architecture of video streaming with its components in detail. The existing of the video streaming service over the internet is another topic that included in this chapter. It included the top 5 metrics such as bit rate, buffer fill, lag length, play length, and lag ratio that measure the performance of video streaming. The fundamental problem that hinders the quality of the video streaming service with its solution also discusses. Finally, what it looks like the services quality monitoring and analysis in Ethio telecom with the key quality indicator (parameters) are described.

Chapter Four
Time Series Data Traffic Forecasting

This chapter introduces the time series and different types of forecasting models and their components. In the forecasting models, it classifies into three categories, such as statistics forecasting model, machine learning models, and deep learning models. In these different categories, it uses some models as an example and describes how it works.

4.1 Definitions of a Time Series

A time series is a sequence of numerical data points listed in time order. Mathematically defined time series as a set of vectors t= 1, 2, 3, ... Where t represents the time elapsed [36]. Most commonly, the data points are distributed evenly in time, with equal spacing between successive data points over the entire time series. Many kinds of data can be gathered into time series, as long as there is a time dependence in the data. Studied time series for several purposes, such as the forecasting of the future based on knowledge of the past, the understanding of the phenomenon underlying the measures, or simply a succinct description of the salient features of the series [37].

A time series containing records of a single variable is called univariate. However, if records of more than one variable are considered, it is termed as multivariate. A time series can be continuous or discrete. In continuous-time, series observations are measured at every instance of time, whereas a discrete-time series contains observations measured at discrete points of time.

A time series can be a stochastic process or a deterministic one. To predict a time series is necessary to use mathematical models that truly represent the statistical characteristic of the sampled traffic. For adaptive applications that require real-time processing, the choice of the prediction method must take into account the prediction horizon, computational cost, prediction error, and the response time.

4.2 Components of Time Series Analysis

A time series, in general, is supposed to be affected by four main components, which can be separated from the observed data. These components are Trend, Cyclical Variations, Seasonal Variations, and Irregular Movements[36]. Figure 4.1 shows the component of time series data. A brief description of these four components is as follows.

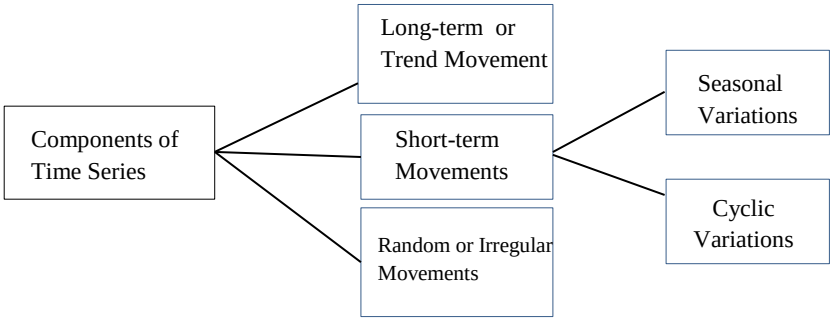

Figure 4.1: Component of time series analysis

4.2.1 Trend

The trend shows the general tendency of the data to increase or decrease during a long period. A trend is a smooth, generally, long-term, average tendency. It is not always necessary that the increase or decrease is in the same direction throughout the given period of time. It is observable that the tendencies may increase, decrease, or are stable in different sections of time. But the overall trend must be upward, downward, or stable.

4.2.2 Seasonal Variations

Seasonality occurs when the time series exhibits regular fluctuations during the same month (or months) every year, or the same quarter every year. These are the rhythmic forces, which operate regularly and periodically for less than a year. They have the same or almost the same pattern for 12 months. This variation will be present in a time series if the recorded data is hourly, daily, weekly, quarterly, or monthly.

These variations come into play either because of the natural forces or man-made conventions. The various seasons or climatic conditions play an important role in seasonal changes. Such as the production of crops depends on the seasons, the sale of umbrella and raincoats in the rainy

season, and the sale of electric fans and A.C. shoots up in summer seasons. The effect of man-made conventions such as some festivals, customs, habits, fashions, and some occasions like marriage is easily noticeable. They recur themselves year after year. Seasonal variation is an essential factor for businessmen, shopkeepers, and producers for making proper plans.

4.2.3 Cyclic Variations

The variations in a time series that operate themselves more than one year are the cyclic variations. This oscillatory movement has a period of oscillation of more than a year. One complete period is a cycle. It is a four-phase cycle comprising of the phases of prosperity, recession, depression, and recovery. The cyclic variation may be regular are not periodic. The upswings and downswings in business depend on the joint nature of economic forces and interaction between them.

4.2.4 Random or Irregular Movements

This component is unpredictable. Every time series has some unpredictable element that makes it a random variable. In prediction, the objective is to "model" all the components to the point that the only component that remains unexplained is the random component. These fluctuations are unforeseen, uncontrollable, unpredictable, and are erratic. These forces are earthquakes, wars, floods, famines, and any other disasters. There is no defined statistical technique for measuring random fluctuations in a time series.

By considering the effects of these four components, it can use two different types of models for a time series viz - Multiplicative and Additive models equation that is shown in equation (1) and equation (2) respectively.

$$Y(t) = T(t)\ X\ S(t)\ X\ C(t)\ X\ I(t) \tag{1}$$

$$Y(t) = T(t) + S(t) + C(t) + I(t) \tag{2}$$

Here $Y(t)$ is the observation and $T(t)$, $S(t)$, $C(t)$, and $I(t)$ are the trends, seasonal, cyclical, and irregular variation at time t respectively. For the multiplicative model, it is work based on the assumption that the four components of a time series are not necessarily independent, and they

can affect one another, whereas, in the additive model, assumed that the four elements/components are independent of each other.

4.3 Forecasting

Forecasting is the act of predicting the future value based on history. A forecasting method is a procedure for computing forecasts from present and past values. Forecasting methods or algorithms used to predict previous data of a time series to get future trends of these data. There are different types of forecasting models, such as statistical forecasting models, machine learning models, and deep learning models**.**

4.3.1 Statistical Forecasting Models

In statistics forecasting, models can be chosen as linear or non-linear depending on whether the current value of the series is a linear or non-linear function of past observations. In general, models for time series data can have many forms and represent different stochastic processes. AR and MA are widely recognized statistical forecasting models that predict future observations of a time series based on some linear function of past values and white noise terms. However, one of the known limitations of such techniques is the poor robustness of the rapid fluctuations of the time-series. Additionally, these methods work with homogeneous time-series, where the input and the prediction are within the same set of values[38].

This section covers some common algorithms used in time series forecasting in the statistical model. For a long time, the forecasting field influenced by linear statistical methods. The AR model, the MA model, and hybrid (ARMA, ARIMA and SARIMA) models that derive from them are an example of linear models.

4.3.1.1 AR model

In the autoregressive process, an output variable y_t depends linearly on its own previous values $\{y_{t-1}, \ldots, y_{t-p}\}$, and some white noise ε_t. By definition, a process $\{y_t\}$ is said to be an autoregressive process of order p denoted AR (p) if y can be described by equation (3).

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_p y_{t-p} + \varepsilon_t \quad (3)$$

Where ε_t is white noise with mean zero and fixed finite variance σ_z^2, and $\alpha_1, \ldots, \alpha_p$ are the parameters of the model. The order p of the model determines the number of past observations used to predict the current value. The simplest example of an AR process is the first-order case, denoted AR (1), given by equation (4).

$$y_t = \alpha_1 y_{t-1} + \varepsilon_t \tag{4}$$

In the multivariable case where there are multiple observations for each time step, then we can consider a multivariate autoregressive or a vector autoregressive (VAR) model. Consider *M* time series generated from *M* variables, a VAR (*p*) model is defined by the following equation (5).

$$y_t = \sum_{k=1}^{p} A_k y_{t-k} + \varepsilon_t \tag{5}$$

Where $y_t = [y_t^{(1)}, y_t^{(2)}, \ldots y_t^{(M)}]^T$ is the M-dimensional time series column vector at index t. Each A_k is an M-by-M matrix of parameters where $A_{i,j}^{(k)}$ is the element at position (i, j) in matrix A_k and $\varepsilon_t = [\varepsilon_t^{(1)}, \varepsilon_t^{(2)}, \ldots \varepsilon_t^{(M)}]^T$ is a column vector of white noises.

4.3.1.2 MA model

Suppose that $\{\varepsilon t\}$ is a purely random process with mean zero and variance δ_2^z, then a process $\{y_t\}$ is said to be a moving average process $\{y_t\}$ of order q denoted MA(q) if y_t can be expressed by equation (6).

$$y_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \ldots + \beta_q \varepsilon_{t-q} \tag{6}$$

Where $\beta_1, \beta_2, \ldots \beta_q$ are parameters of the model

The moving average also describes a method where the next sample depends on the weighted sum of the past or present inputs of an exogenous time series $\{x\}$ of dimensions described in the equation (7) below.

$$y_t = \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \ldots + \beta_q x_{t-q} \tag{7}$$

Similar to the AR (p) model, in the case of multiple time series, a multivariate MA (q) model of M dimension can be written as equation (8).

$$y_t = \sum_{k=0}^{q} B_k y_{t-k} \quad (8)$$

Where x_t is an exogenous N-dimension time series, and B_k are M-by-N matrices of parameters.

4.3.1.3 ARIMA models

ARIMA models are well-known for their essential forecasting accuracy and flexibility in representing several different types of time series. It introduces a differencing process that effectively transforms the non-stationary data into a stationary one. This is done by subtracting the observation in the current period from the previous one. But, a significant limitation is their presumed linear form of the associated data that makes them inappropriate for sophisticated nonlinear time series modeling. Hence, the ARIMA model is called "Integrated" ARMA. ARIMA model aims to describe the autocorrelations in the data. The general form of the ARIMA (p, d, q) process described in equation (9).

$$y'_t = \nabla^k y_t = \alpha_1 y'_{t-1} + \cdots + \alpha_p y'_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q} \quad (9)$$

Where parameters *p*, *d*, and *q* are non-negative integers that refer to the order of the autoregressive part, the degree of first differencing involved, and the order of moving average part, respectively.

To determine the proper model for the given time series, determine the values of p, d, and q, it is necessary to carry out Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) analysis. These statistical measures reflect how the observations in a time series are related to each other.

4.3.1.4 SARIMA models

For time-series data with the seasonality components present, it is possible to consider an extension of ARIMA model known as SARIMA This model has non-seasonal and seasonal parts, both having the AR, Integrated/ difference and MA parameters. It used when the data present with a periodic characteristic which must be known ahead. For example, the seasonal component that repeats every s observations can be monthly s = 12 (12 in 1year) or quarterly s = 4 (4 in 1 year). The SARIMA model usually termed as ARIMA (p, d, q) X (P D Q)s where

P=number of seasonal autoregressive (SAR) terms, D=number of seasonal differences, Q=number of seasonal moving average (SMA) terms[39].

The order of the seasonal or non-seasonal AR and MA components determined by the ACF and PCF plots peak values behavior (i.e., decay to zero) of the lags.

4.3.2 Machine Learning Forecasting Models

Machine learning (ML) is a data analytics technique that teaches computers to do what comes naturally to humans and animals: learn from experience. ML algorithms use computational methods to "learn" information directly from data without relying on a predetermined equation as a model. It finds natural patterns in data that generate insight and help you make better decisions and predictions. The algorithms adaptively improve their performance as the number of samples available for learning increases.

Machine learning uses two types of techniques, such as supervised learning and unsupervised learning. When training a supervised learning algorithm, the training data will consist of inputs paired with the correct outputs. During training, the algorithm will search for patterns in the data that correlate with the desired outputs. Unsupervised learning to allow the model to work on its own to discover information. It mainly deals with the unlabelled data.

ML approach can analyze features, relationships, and complex interactions among features of a problem from samples of a dataset and learn a model, which can be used for demand forecasting [40]. In ML, the main advantage of forecasting is to develop automated algorithms that can learn and make some predictions from data. The difference lies in how such minimization of error is done with ML methods utilizing nonlinear algorithms to do so while statistical once linear processes. It used commonly based on ANNs.

Suggested ANNs approach as an alternative technique to time series forecasting, and it gained immense popularity in the last few years [41]. Similar to the work of a human brain, ANNs try to recognize regularities and patterns in the input data, learn from experience, and then provide generalized results based on their known previous knowledge. Although the development of ANNs was mainly biologically motivated, afterward, they applied in many different areas, especially for forecasting and classification purposes.

The standard type of ANNs consists of three layers of units/nodes. The input layer of nodes feeds the input variables into the network, and the hidden layer is a bridge between the input layer and the output layer. It was adjusting the weights during the training process. The nodes in each layer (excluding the input layer) will compute a weighted sum of the inputs and perform a nonlinear transformation (functional mapping) on the sum using different activation functions. A specific ANN determined by its topology, learning paradigm, and learning algorithm. Figure 4.2 represented a simple neural network.

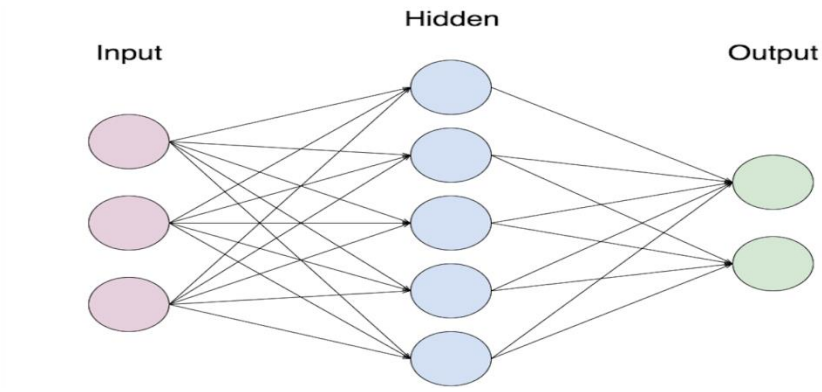

Figure 4.2: Artificial Neural Network topology

4.3.2.1 Multilayer Perceptron

The MLP is the popular and most frequently used type of neural network. It is composed of more than one perceptron. They are composed of an input layer to receive the signal, an output layer that makes a decision or prediction about the input, and in between those two, an arbitrary number of hidden layers that are the right computational engine of the MLP. MLPs with one hidden layer is capable of approximating any continuous function.

On most occasions, transmit the signal within the network in one direction: from input to output. There is no loop; the output of each neuron does not affect the neuron itself. This architecture is called feedforward. Figure 4.3 represents the network diagram for a MLP with two layers of weights.

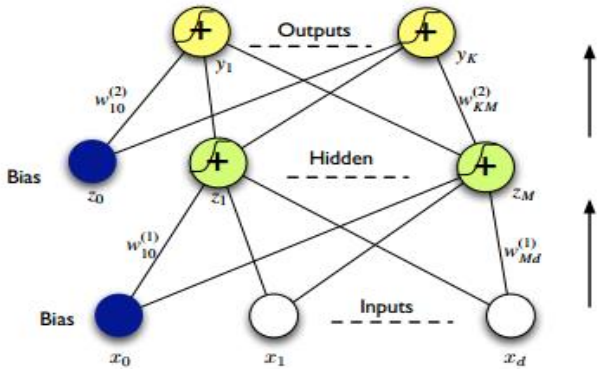

Figure 4.3: Network diagram for a multilayer perceptron (MLP)[42]

4.3.2.2 Extreme Learning Machine

ELM is an ANN with a unified learning platform developed to improve the efficiency for single layered Feedforward Networks. ELM theory show that the value of the weight of this hidden layer need not to be tuned, and be therefore independent of the training data. EML can solve any regression problem with a desired accuracy, if it has enough hidden neurons and training data to learn parameters for all the hidden neurons. Figure 4.4 shows the Feedforward network architecture to implement ELM algorithm.

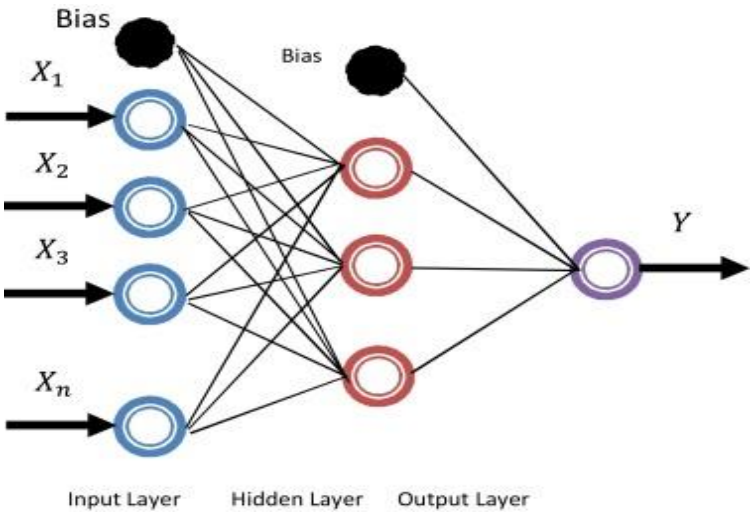

Figure 4.4: Feedforward network architecture to implement ELM algorithm

4.3.3 Deep Learning Forecasting Models

Deep learning (DL) is a ML technique that applies deep neural network architectures to solve various complex problems. DL is an implementation of artificial neural networks, which mimic the natural human brain. However, a deep neural network is more powerful and capable of analyzing and composing more complex features and relationships than a traditional neural network. DL requires high computing power and large amounts of data for training. The recent improvements in the Graphical Processing Unit (GPU) and parallel architectures enabled the necessary computing power required in deep neural networks. DL uses successive layers of neurons, where each layer extracts more complex and abstracts features from the output of previous layers. Thus, a DL can automatically perform feature extraction in itself without any preprocessing step. Visual object recognition, speech recognition, and genomics are some of the fields where DL is applied successfully. DL has four network architectures, such as Unsupervised Pre-trained Networks, Convolutional Neural Network, A Recurrent Neural Network (RNN), and A Recursive Neural Network. From these four architecture of DL, we select the RNN architecture because the RNN uses for time steps. So we can discuss it in detail in the next section.

4.3.3.1 Recurrent Neural Network

The RNN is a class of neural network whose connections of units form a directed cycle; this nature can work with temporal data. The recurrent is defined as there exists a path of one or more cycles from a unit back to itself [43]. It introduces multiple hidden layers; there are connections between hidden layers to hidden layers, which is the most common architecture of RNN as shown in Figure 4.5, the expression is denoted in equation (10) and (11) as follows as.

$$h^t = \delta(W_h X_t + W_r\, h^{t-1}) \tag{10}$$

$$y_t = \delta(W_y h^t) \tag{11}$$

Where X_t denotes input at time t,

W_h denotes weight matrix from the input layer to the hidden layer

W_r denotes the weight of recurrent computation,

W_y denotes weight from hidden layer to the output layer,

h^t denotes values of hidden nodes at the time of t,

y_t denotes a value of output node at the time of t.

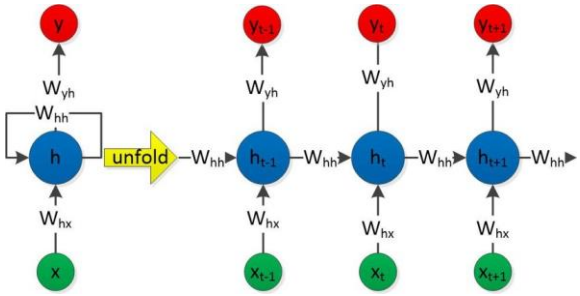

Figure 4.5: unfold Architecture of RNN

RNNs are capable of learning features and long term dependencies from sequential and time-series data. ANNs with recurrent connections are called RNNs, which are capable of modeling sequential data for sequence recognition and prediction. RNNs are also a feed-forward network, however with recurrent memory loops that take the input from the previous or same layers or states. This memory gives them a unique capability to model along the time dimension and arbitrary sequence of events and inputs data. One of the most common types of RNN models is LSTM network.

4.3.3.2 Long Short-term Memory Network

LSTM networks are a special kind of RNN that is capable of learning long-term dependencies. In regular RNN, small weights are multiplied over and over through several time steps, and the gradients diminish asymptotically to zero- a condition is known as vanishing gradient problem. So LSTM designed to avoid the long-term dependency issue, which is the cause of the vanishing gradient problem in normal RNNs. The major innovation of the LSTM is its special unit called Memory Block or LSTM unit, which has self-connections storing the temporal state of the network [44].

As indicated in Figure 4.6, the architecture of the LSTM Recurrent Neural Network is like ANN. However, hidden units of LSTM contain connections, which is the general architecture of RNN; besides, each hidden unit used for memorizing and forgetting the goal.

LSTM network typically consists of memory blocks, referred to as cells, connected through layers. The information in the cells contained in cell state C_t and hidden state ht, and it regulated by mechanisms, known as gates, through sigmoid and tanh activation functions. LSTM, therefore, can, conditionally, add or delete information from the cell state. In general, the gates take in, as input, the hidden states from previous time step h_{t-1}, and the current input x_t and multiply them pointwise by weight matrices, W, and a bias b added to the product.

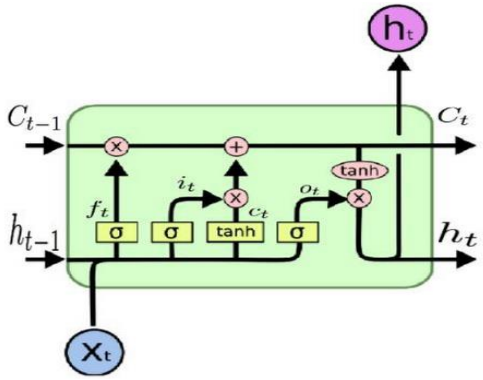

Figure 4.6: Architecture of LSTM models

Tanh activation uses to help regulate the values flowing through the network. The tanh function squishes values to always be between -1 and 1.

A sigmoid activation is similar to the tanh activation. Gates contains sigmoid activations. Instead of squishing values between -1 and 1, it squishes values between 0 and 1. 0 indicating *Nothing* goes through and 1 implying *Everything* goes through. Figure 4.7 shows the graphical representation of the sigmoid versus tanh function.

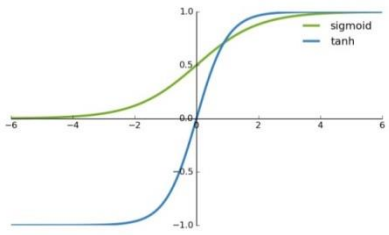

Figure 4.7: Sigmoid versus than function

An LSTM unit is composed of three gates and one cell, such as an input gate, an output gate, a forget gate, and a memory cell, which used to control the flow of information. Memory cell essentially acts as an accumulator of the state information; the cell accessed, written, and cleared by several self-parameterized controlling gates. Every time a new input comes, its data will be accumulated to the cell if the input gate is activated, and the prior cell status will be forgotten in the process if the forget gate is on and enabled. The final state will propagate to the final state controlled by the output gate. Overcoming the vanishing too quickly from the RNN model is a critical improvement for LSTM. The multivariate version of LSTM is where the input, cell output, and states are all vectors. The three main gates are:

- Forget gate:

After getting the output of the previous state, $\mathbf{h_{t-1}}$, Forget gate helps us to make decisions about what must be removed from $\mathbf{h_{t-1}}$ state and thus keeping only relevant stuff. It is surrounded by a sigmoid function, which allows crushing the input between [0,1] with 0 meaning delete all and 1 implying remember all. It represented in equation (12).

$$f_t = \delta(W_{xf}x_t + U_{hf}h_{t-1} + V_{cf} \cdot c_{t-1} + b_f) \quad (12)$$

- Input gate:

In the input gate, we decide to add new stuff from the present input to our current cell state scaled by how much we wish to add them. The sigmoid layer determines which values to be updated, and the tanh layer creates a vector for new candidates to add to the present cell state. Equation (13) and equation (14) reperesent the input and new candidates cell respectively.

$$i_t = \delta(W_{xi}x_t + U_{hi}h_{t-1} + V_{ci} \cdot c_{t-1} + b_i) \quad (13)$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot tanh(W_{xc}x_t + U_{hc}h_{t-1} + b_i) \quad (14)$$

- Output gate:

Finally, for deciding what to output from our cell state, which will be done by our sigmoid function. By multiplying the input with tanh to crush the values between (-1,1) and then multiply it with the output of sigmoid function so that we only output what we want to. Equation (15) and equation (16) reperesent the final out put.

$$o_t = \delta(W_{xo}x_t + U_{ho}h_{t-1} + V_{co} \cdot c_{t-1} + b_{io} \quad (15)$$

$$h_t = o_i \cdot \tanh(c_t) \quad (16)$$

There are several models to forecast the time series proposed in the literature for network traffic. These can be classified into two categories: linear prediction and nonlinear prediction. The most widely used traditional linear prediction methods are ARIMA, FARIMA, and SARIMA model [11], [12], [15], and [18]. The most common nonlinear forecasting methods involve neural networks (NN) are: [13], [14], [16], and [17]. The experimental results from [19] show that nonlinear traffic prediction based on NNs outperforms linear forecasting models (SARIMA-ELM and SARIMA-MPL), which cannot meet the accuracy requirements. So we need to find the deep learning models that forecast the video streaming data traffic that capture the character of video streaming, i.e., self-similarity and long term dependence.

4.4 Chapter Summary

In this chapter, to discuss the time series model that forecasting the network traffic. Three models to predict the network traffic, such as the statistical forecasting model, machine learning forecasting model, and deep learning models. In the statistical model, there are only capturing the linear part of the network traffic characteristics and ignore the nonlinear elements. However, the limitations of these techniques are poor robustness to the rapid fluctuations of time-series. Additionally, these methods work with homogeneous time-series, where the input and the prediction are within the same set of values. ANNs successfully overcome the drawback of statistical models. The excellent feature of ANNs is their inherent capability of nonlinear modeling, without any presumption about the statistical distribution followed by the observations. One of the limitations of the neural network is that there is no memory associated with the model.

In the deep learning forecasting model, RNNs are one of the architectures that done on the time steps. RNNs called recurrent because they perform the same task for every element of a sequence, with the output depended on the previous computations. Form this architecture LSTM model is the one that handles both self-similarity and long term dependence character in addition to the traditional NN forecasting models.

Chapter Five
Experimental Analysis and Results

In this chapter, the overall experimental process, followed by the result, is described in detail. In the system model, we will see data collection and data preprocessing techniques. Finally, evaluate the proposed forecasting model and discussed it with other hybrid forecasting models.

5.1 System Model

In the forecasting principle, the forecasting model built is expected to capture all characteristics and components of the predictor data set. The system model designed in the thesis divides the data set into training and test data for the LSTM, SARIMA_ELM, and SARIMA_MPL models. In the training set, the first 80 percent of video streaming data traffic starting from October 2018, is used to investigate efficient parameters and specification of the model. While in the test set, the last 20 percent of the video streaming data traffic, is reserved for the evaluation of out-of-sample and performance comparison among prediction models. The models are evaluated based on different performance metrics. The system model to perform the modeling and forecast is illustrated in Figure 5.1.

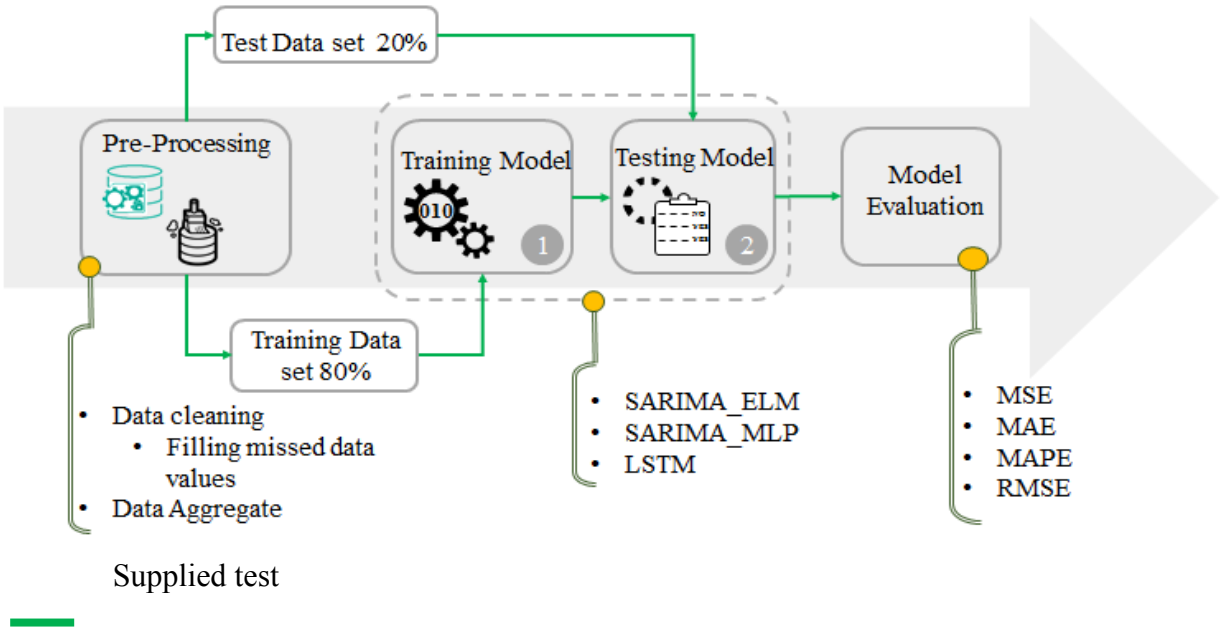

Figure 5.1 The system model flaw chart

5.2 Data Set

In this thesis, we considered a daily-based measurement of the video streaming data traffic network generated from the platform on Huawei Smart Care Solution for the period from October 2018 to July 2019. In this study, one historical value i.e., the total volume of video streaming data traffic, is used as an input variable. The output of the prediction model is the total volume of video streaming data traffic in the next two months. To perform the data analysis, model building, and forecasting, an open-source and statistical software called SPSS and R software with its builtin and modified packages have used.

5.3 Data Preprocessing

We have aggregated the data set collected on an hourly basis from five RNCs daily. The data cover 304 days. During data analysis, we were able to get missing data for June 2019 and July 2019 , as shown in Figure 5.2 (circled with red color). From a total of 304 days of data, 15 days of data value is missing, as indicated in Figure 5.3. This hole happened because of service interruption in the country. We use SPSS software, i.e., series means algorithm estimation was used for smoothing the series as well as estimate the missed values, therefore to keep consistency and completeness of the data for analysis.

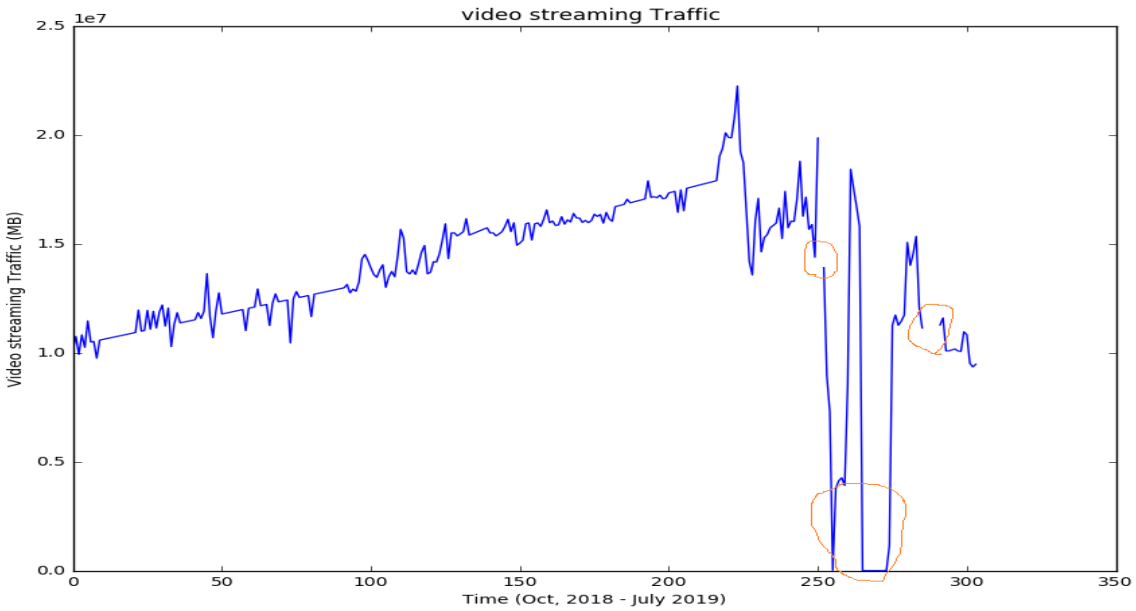

Figure 5.2: UMTS video streaming data traffic collected daily

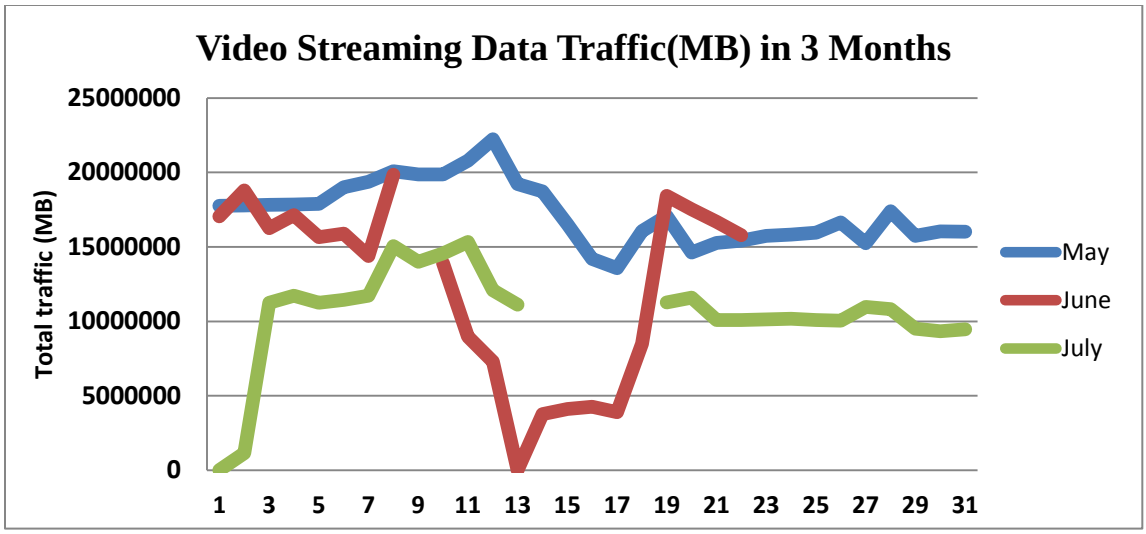

Figure 5.3: Three months (May 2019 – July 2019) video Streaming Data Traffic

Figure 5.4 represent the total video streaming data traffic after the missed data was handled.

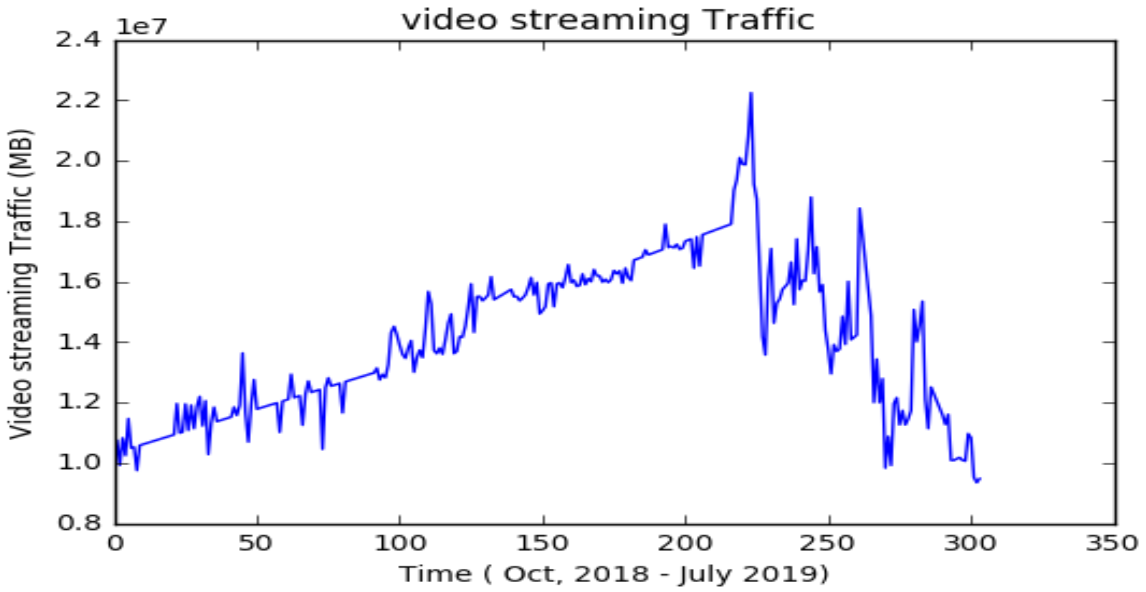

Figure 5.4: UMTS Video Streaming data traffic in daily-basis (October 2018 –July 2019)

5.4 Training and Test Data

Divide the entire dataset into two. Those were training data sets (the first 80% of the whole data set) and test data sets (the remaining 20%), as indicated in section 5.1. Figure 5.5 shows the classification of the data into training and test data.

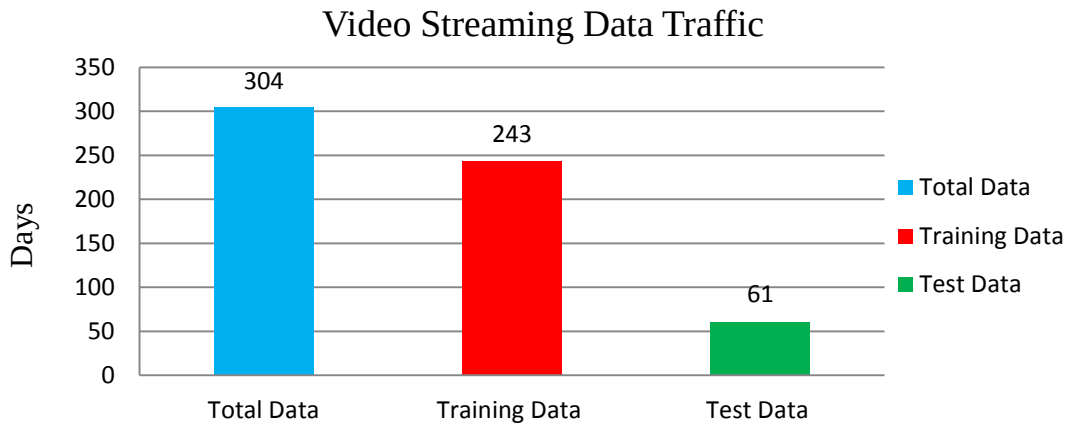

Figure 5.5: Training and Test data

5.5 Daily, Weekly and Monthly Trends of Video Streaming Data Traffic

Figure 5.6 shows the daily trend of video streaming data traffic in hourly-basis. After retrieving the entire data, it is possible to start structuring the raw data set, into the different data collections. As indicated in the graph, within 24 hours, start the usage early in the morning around 6:00 and continue until 22:00. Until 18:00, the usage increases at a decreasing rate. Afterward, the usage of the data traffic increases at an increasing rate until midnight. It also indicates, after working hours, video streaming usage increases. This trend is the same for every day. The other thing that we see from this graph is the weekend days is a higher usage when we compare to the weekday traffic.

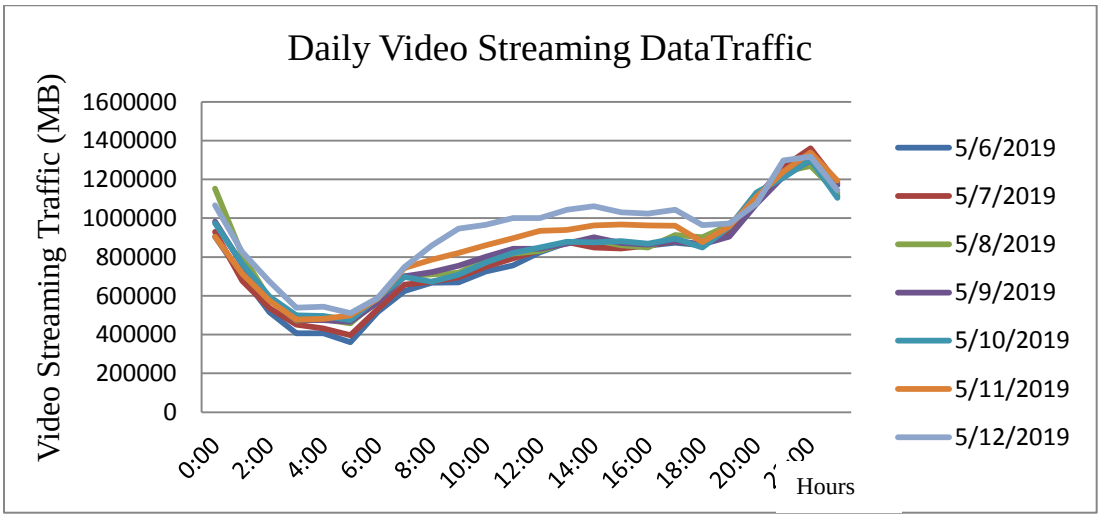

Figure 5.6: Daily Usage of Video Streaming DataTraffic from May 6-12, 2019

Figure 5.7 depicts the total video streaming data traffic for the six weeks observation period from August 29, 2018, to October 8, 2018. The graph shows the difference in behavior and traffic load between weekdays and weekends.

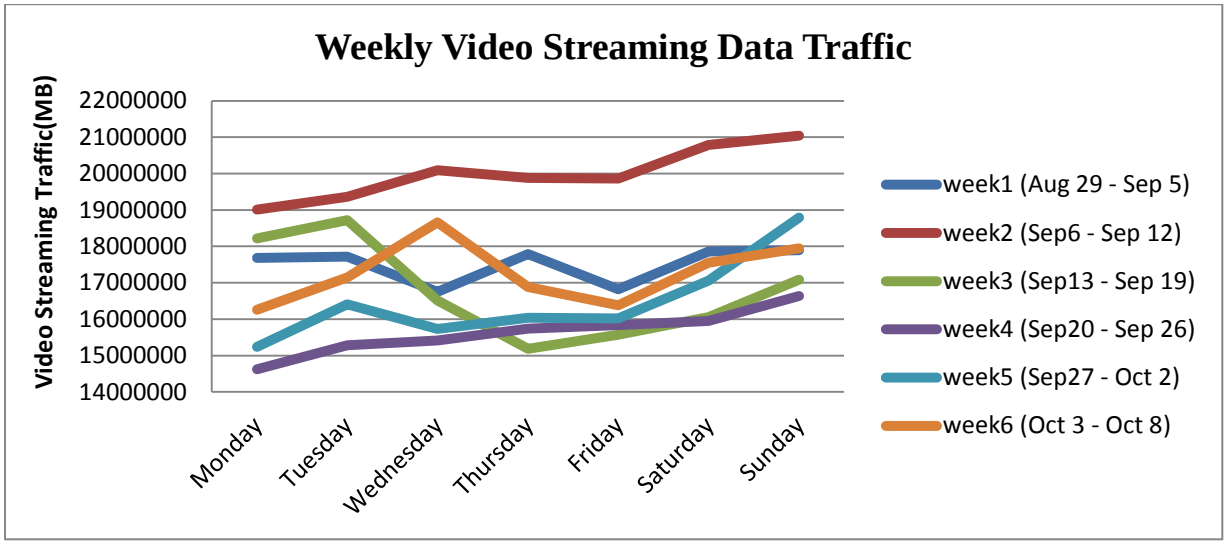

Figure 5.7: Traffic usage data observations in weekly basis

Figure 5.**8** shows the monthly usage pattern of video streaming data traffic. As indicated in the graph, the video streaming traffic is increasing at an increasing rate from October 2018 to May 2019. In June 2019 and July 2019, there is an irregular pattern (ups and down) because of service interruption.

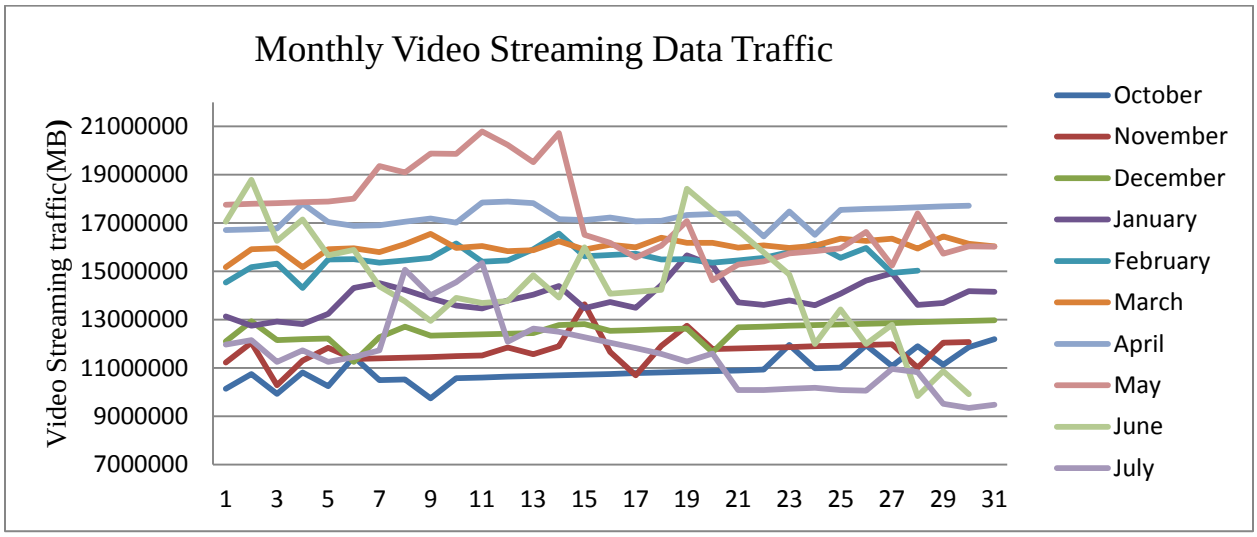

Figure 5.8: Monthly Distribution of video streaming data traffic

5.6 Component Identification

Figure 5.9 illustrated the decomposition of the time series. It examined for features such as trend and seasonality. The pattern across those time units indicates a seasonal pattern. The observed regularly repeating pattern of highs and lows relates to seasonal data of months of the year, which show seasonality.

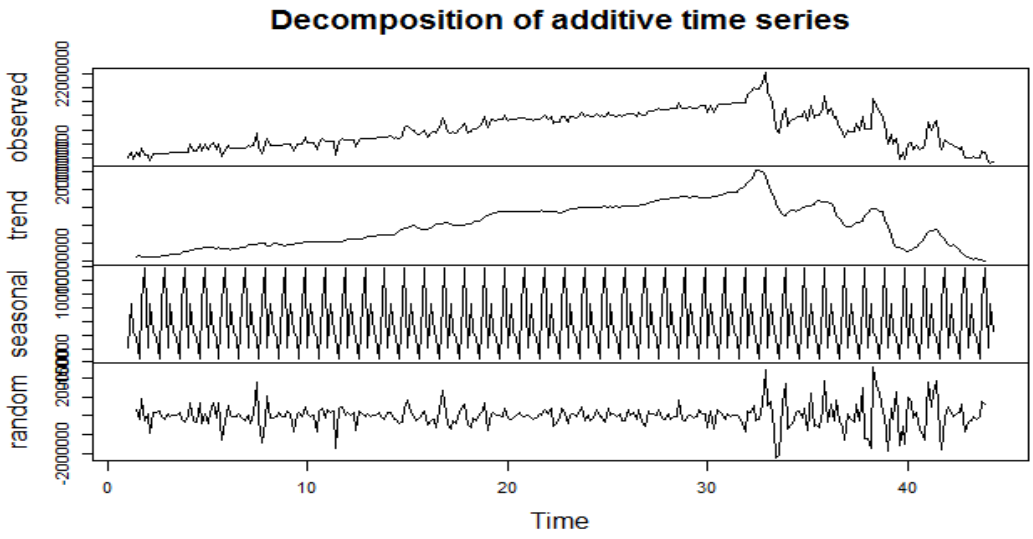

Figure 5.9: Decomposition of the UMTS data traffic into time series components

5.7 Model Selection and Fit

Based on the minimum values of AIC, BIC, and AICc as shown in Figure 5.10, SARIMA (0,1,1) $(1,1,1)_7$ is selected as the best fit linear model for the UMTS video streaming data traffic (see Annex). From the fitting plot of the selected SARIMA **(**0,1,1) $(1,1,1)_7$ model shown in Figure 5.11, the spikes out of the curve fit indicates the existence of non-linearity on the UMTS peak data traffic. This non-linearity to be extracted as a residual or an error, and modeled with MLP and ELM.

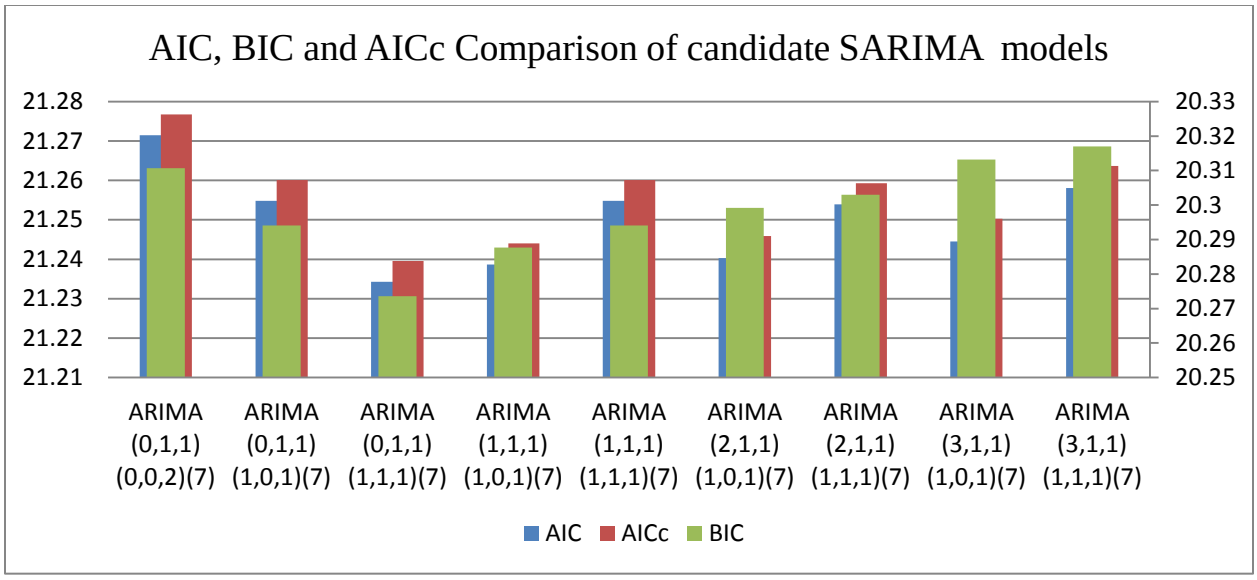

Figure 5.10 Candidate SARIMA models comparison for selection

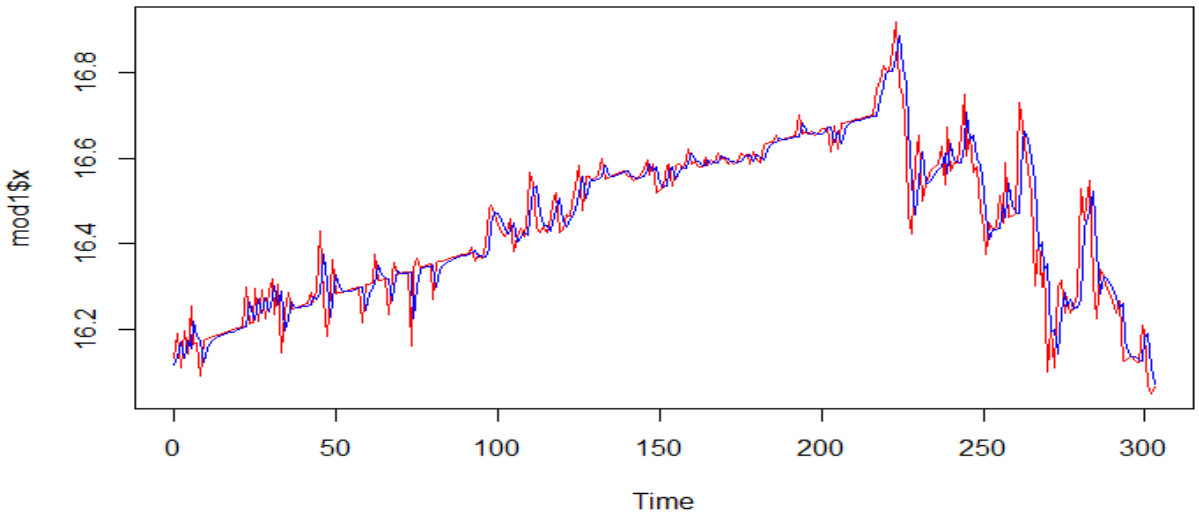

Figure 5.11: SARIMA $(0,1,1)(1,1,1)_7$ model fit to the UMTS data

Diagnosis of model fit from its residuals

In Figure 5. **12** below, the ACF plots of the residuals lag values lie within a p-value of 0.1 and -0.1. The Q-Q plot shows the residuals distributed normally and uncorrelated, which is informative that the selected model fitted well.

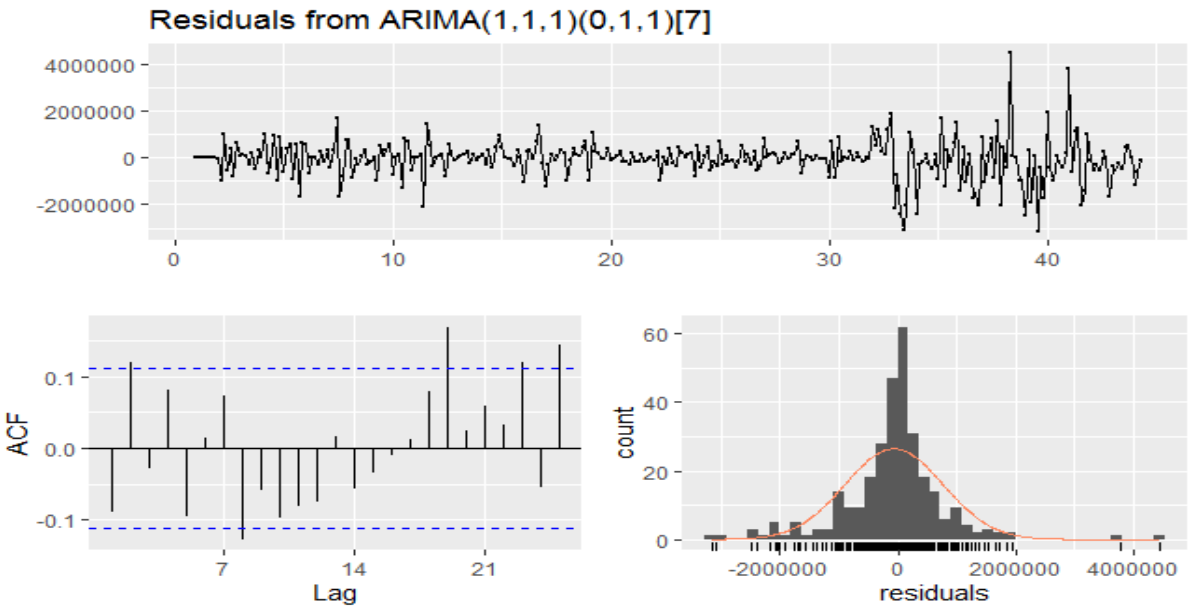

Figure 5. 12 Diagnosis of SARIMA $(0,1,1)(1,1,1)_7$ model

5.8 Numerical Results

5.8.1 Evaluation Setup

We use the set of video streaming data traffic from five RNCs that collected during ten months, to evaluate the performance of the proposed model. For each RNCs, we calculate the aggregate cell traffic, from hourly-bases to daily-bases, as described in Section 5.3.

Here it defines the performance metrics used to compare the performance of the different models developed in the thesis for the measurement of the accuracy of the prediction algorithm. There are MAE, MSE, MAPE, and RMSE. Models with a minimum of these statistics are considered the best for forecasting. For output value y_i, it is defined in equation (17), (18) , (19) and (20) as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (17)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(|y_i - \hat{y}_i|)^2} \qquad (18)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{n}\right| \qquad (19)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (20)$$

The prediction algorithm implementation of the video streaming data traffic is done in Python, using Keras and Tensorflow, as backend and R statistical software. In addition to deep learning models, non-deep learning methods, i.e., SARIMA, is also implemented for comparison. SARIMA is a commonly utilized model for time series analysis.

Table 5. 1 indicate the chosen hyperparameters that used for the LSTM model. The number of hidden layers is fixed to 5: this is one of the hyperparameters that need to be selected and can affect the tradeoff between the prediction accuracy and the time required to train the network. A higher number of layers may increase the precision of the prediction. Still, we want to focus on the relationship between the number of past observed values and the accuracy of the multi-step prediction, which determines the quantity of information needed to be memorized and utilized by the network. For the same reason, we fix the number of epochs to 100. We use eight months of data for training and two months of data to test/validate the model. We use Adam optimization to update the network weights iteratively based on the training data.

Table 5. 1 Training Hyperparameters

No	Parameter	Remark
1	Initial Learning Rate	0.001
2	LSTM Hidden States	64
3	LSTM Hidden Layers	5
4	Feedforward Hidden Layers	1
5	Num. of Epochs	100
6	Optimization Algorithm	Adam
7	Loss Function	MAE

5.8.2 Results Analysis

This thesis aims to forecast the video streaming data traffic based on past observation and make the prediction an input for planning and capacity analysis. The result of our analyzer based on the total video streaming data traffic as input parameters. We present the results of multi-step

prediction (N=61); that is when the output delayed for a fixed number of timeslots and the forecast performed for later time instants.

In Figure 5.11 and Figure 5.**13** shows the results of the video streaming traffic fitting plot for the selected two different models (SARIMA and LSTM) since they are using the same data and tools. We can see that the prediction is precise for the whole months, despite the oscillating behavior of the traffic.

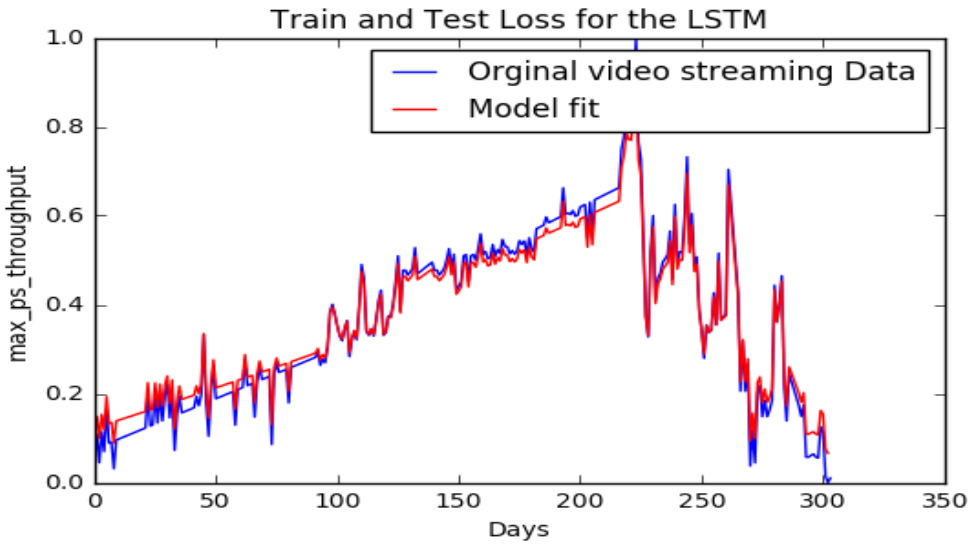

Figure 5.13: LSTM model fit to the UMTS data

From the complete data set, 304 days of video streaming data traffic, 243 days, or eight months data uses for the training purpose. The remaining 61 days, we use for validation of the model for all models. In Figure 5.**14**, it shows the UMTS data traffic was forecasted for two months ahead with SARIMA $(0,1,1)(1,1,1)_7$ model. Based on Figure 5.**15**, the LSTM model to predict the test data better than the other two models (SARIMA-ELM and SARIMA-MLP models). This better prediction is because the LSTM model included self-similarity and long-term dependency character of the video streaming data traffic that not included in the SARIMA-ELM and SARIMA-MLP models.

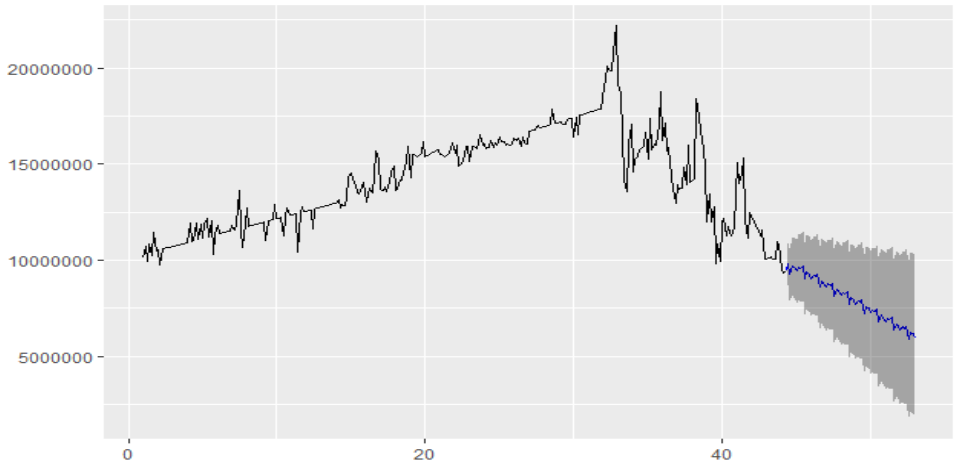

Figure 5.14: SARIMA $(0,1,1)(1,1,1)_7$ model test set prediction

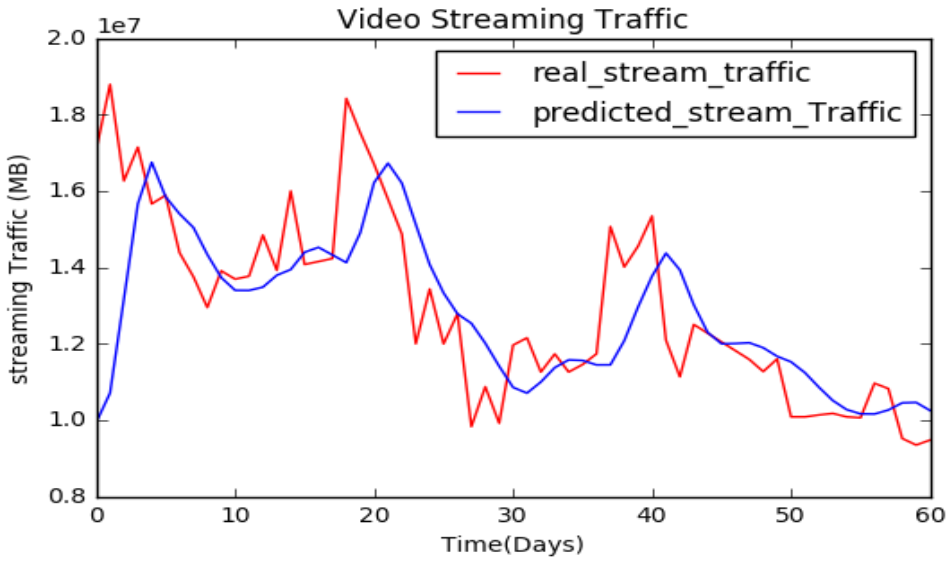

Figure 5.15: LSTM model test set prediction.

As shown in Figure 5.**16**, the performance evaluation of the three models is presented. The Figure indicates that the deep learning, LSTM, forecasting models have a minimum forecasting error compared to the hybrid models, i.e., the statistical and ANN models (SARIMA-ELM, SARIMA-MPL).

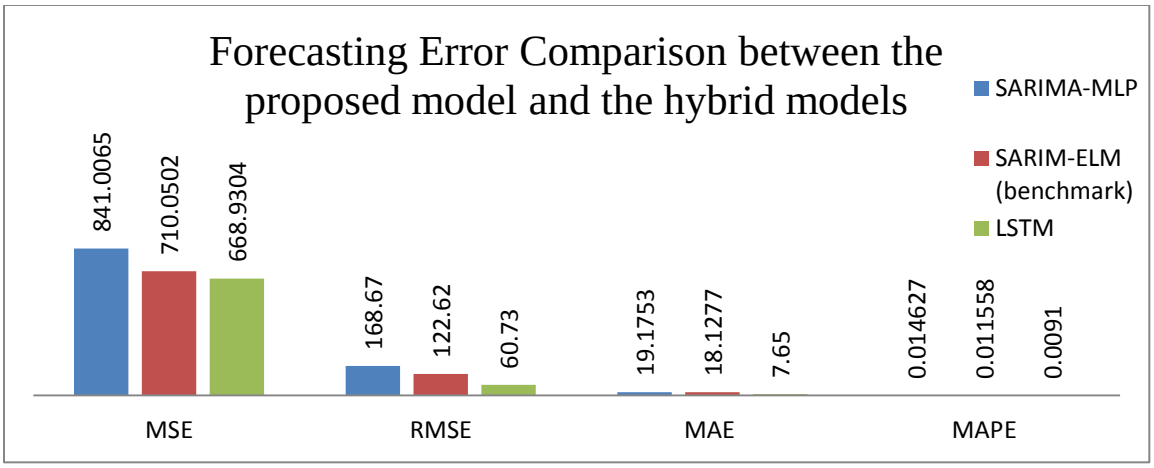

Figure 5.16: Forecasting error comparison between the LSTM and the hybrid models

5.9 Discussion

There are many statistical forecasting time series models that predict future value based on past observation. However, this model works only the linear part of the traffic and homogeneous time series. The machine learning model incorporates the nonlinear elements of the data traffic but did not work for long term dependence and self-similarity characteristics. Among deep learning models, the LSTM model used for the prediction of video streaming data traffic because it handles the character of the video streaming data traffic such as long term dependency and self-similarity, which is not incorporated by the other models.

In this study, as indicated in Figure 5.16, the LSTM model presents better performance than the SARIMA-ELM and SARIMA-MLP models in all error measures. The SARIMA-ELM model was used as a benchmark for this work because this model was proven by the previous (last year) work, which is done on mobile data traffic prediction in ethio telecom[19]. The predicted MSE of the benchmark model is 710.0502, while the predicted MSE of the LSTM model is 668.9304, and the prediction result enhanced by 5.79% compared to the benchmark model. When we compare the prediction of MAE in the benchmark model is 18.1277, while the predicted MAE of the proposed model is 7.650, so the prediction result enhanced by 57.8% compared to the benchmark model. Lastly, the predicted MAPE of the benchmark, which expresses accuracy as a percentage of error, is 1.15%. In comparison, the MAPE of the LSTM model is 0.91%, so the prediction result improved by 21.27% compared to the benchmark model. In general, in all performance metrics, the proposed model has minimum values of error.

5.10 Chapter Summary

This chapter discusses the experimental analysis of Ethio telecom video streaming data traffic of ten months. The data collection period of this study starts from October 2018 to July 2019 from five RNCs. The data were preprocessed and categorized into training and test data. The data were entered into the R statistics tools and discuss the result. Based on the analysis, the LSTM model had a minimum forecasting error.

Chapter Six
Conclusion and Recommendation

6.1 Conclusion

The main aim of this thesis is to model and forecast UMTS video streaming data traffic demand based on historical data collected from the ethio telecom network. And also to evaluate three Univariate time series methods, namely SARIMA-ELM, SARIMA-MPL, and LSTM model, to predict the UMTS video streaming data traffic. The intention was to find a model that fits well with the data and could forecast the UMTS data-traffic in contrast to the forecasting performed today.

The prediction of video streaming data traffic is difficult, mainly on whether the model or methodology can capture the self-similarity and long-term dependency. We conclude that the data processing step is significant in generating the results. We constructed a prediction model of video streaming data traffic based on RNN using LSTM units, which is one of the typical methodologies of deep learning. The LSTM model used in this thesis is a deep learning architecture for expressing nonlinear and complex features of the video streaming data traffic more effectively. The LSTM model can learn more information from the past (243 days of video streaming data traffic) and can help to predict (the future 61 days of video streaming data traffic) more reliable video streaming data traffic.

Besides, LSTM is suitable for temporal data prediction; the special recurrent connections between hidden layers to hidden layers are essential for the model to remove the unnecessary information.

The experimental results present, the proposed model has lower RMSE, MAE, MAPE, and MASE error metrics compared with the other models (SARIMA-ELM, SARIMA-MPL). An average of 57.8% of mean absolute error (MAE) improvement is found. The overall result demonstrates that an LSTM approach can be an effective method for predicting/forecasting video streaming data traffic to reflect temporal patterns. So the LSTM model can include more features in video streaming data traffic that is not incorporated by the hybrid models (SARIMA-ELM, SARIMA-MPL).

6.2 Recommendations

Based on the analysis of the available data and conclusions drawn, we forwarded the following recommendations.

- The results obtained from this research will go a long way in helping ethio telecom to use the model to forecast the video streaming data traffic.
- LSTM forecasting model can be used by ethio telecom for planning and resource allocation purposes.
- This study can serve as a guide to telecommunications equipment manufacturing companies to improve on engineering designs of network transmission equipment.

6.3 Future work

In the future, researchers shall consider the UMTS video streaming data traffic as multiple variable data series and including different factors as features like customers' behavior modeling. In addition to this, they shall use a Combining Approach, combine different dissimilar models to improving forecast accuracy.

Reference

[1] V. A. Siris, M. Anagnostopoulou, and D. Dimopoulos, "Improving mobile video streaming with mobility prediction and prefetching in integrated cellular-WiFi networks," *Lect. Notes Inst. Comput. Sci. Soc. Telecommun. Eng. LNICST*, vol. 131, pp. 699–704, 2014.

[2] H. Assem, B. Caglayan, T. S. Buda, and D. O'Sullivan, "St-dennetfus: A new deep learning approach for network demand prediction," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11053 LNAI, pp. 222–237, 2019.

[3] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022 White Paper," 2019. .

[4] R. Pereira and E. G. Pereira, "Video streaming," *Pervasive Comput.*, pp. 417–444, 2016.

[5] D. Gupta, "Multi-Step-Ahead Prediction of MPEG-coded video source Traffic using Empirical Modeling Techniques," 2004.

[6] N. Subhash, "FORECASTING TELECOMMUNICATIONS DATA WITH," vol. 3, no. 5, pp. 239–244, 2015.

[7] R. Mouly and P. Mangnale, "An Assessment Of Ethiopian Telecom Customer Satisfaction," vol. 10, no. 4, pp. 10–15, 2010.

[8] Federal Democratic Republic of Ethiopia, "The Second Growth and Transformation Plan (GTP II) - Midterm Review Report," no. June, 2018.

[9] A. Salem, "Seasonal Auto-Regression Integrated Moving Average-based Data Traffic Forecasting : The Case of UMTS Network in Addis Ababa , Ethiopia," 2016.

[10] B. Zhou, D. He, and Z. Sun, "Traffic modeling and prediction using ARIMA/GARCH model," no. May, 2014.

[11] A. Sang and S. qi Li, "A predictability analysis of network traffic," *Comput. Networks*, vol. 39, no. 4, pp. 329–345, 2002.

[12] B. Zhou, D. He, Z. Sun, and W. H. Ng, "Network traffic modeling and prediction with ARIMA/GARCH," *HET-NETs' 06 Conf.*, no. September, pp. 1–10, 2006.

[13] M. Oravec, M. Petráš, and F. Pilka, "Video traffic prediction using neural networks," *Acta Polytech. Hungarica*, vol. 5, no. 4, pp. 59–78, 2008.

[14] V. B. Dharmadhikari and J. D. Gavade, "An NN approach for MPEG video traffic

prediction," *ICSTE 2010 - 2010 2nd Int. Conf. Softw. Technol. Eng. Proc.*, vol. 1, no. November 2010, 2010.

[15] A. K. Al-Tamimi, R. Jain, and C. So-In, "High-definition video streams analysis, modeling, and prediction," *Adv. Multimed.*, vol. 2012, 2012.

[16] A. Abdennour, "Short-term MPEG-4 video traffic prediction using ANFIS," *Int. J. Netw. Manag.*, vol. 15, no. 6, pp. 377–392, 2005.

[17] A. Khan, L. Sun, and E. Ifeachor, "Content-Based Video Quality Prediction for MPEG4 Video Streaming over Wireless Networks," no. August, 2009.

[18] D. R. Marković, A. M. Gavrovska, and I. S. Reljin, "4K video traffic prediction using seasonal autoregressive modeling," *Telfor J.*, vol. 9, no. 1, pp. 8–13, 2017.

[19] G. Tesfaye, "Hybrid SARIMA-ELM-based Data Traffic Forecasting : The Case of UMTS Network in Addis Ababa , Ethiopia," 2018.

[20] H. Holma, M. Kristensson, J. Salonen, and A. Toskala, *WCDMA for UMTS: HSPA Evolution and LTE: Fourth Edition*. 2008.

[21] L. J. Vora, "Evolution of Mobile Generation Technology: 1G To 5G and Review of Upcoming Wireless Technology 5G," *Int. J. Mod. Trends Eng. Res.*, vol. 2, no. 10, pp. 281–290, 2015.

[22] G. Dereje, "UMTS Traffic Model Using Artificial Neural Network: The Case of Addis Ababa, Ethiopia," 2017.

[23] L. Mitikie, "UMTS Coverage and Capacity Planning for the case of Bole Sub City in Addis Ababa," 2016.

[24] A. Margarida and P. Simões, "Temporal Modelling of Mobile Data Traffic Applications for Network Optimisation," no. May, 2017.

[25] M. S. Peter Schefczik, "Evolution of the UTRAN architecture," *Innovation*, no. July, pp. 4–5, 2003.

[26] P. A. Ochang and P. J. Irving, "Evolutionary Analysis of GSM , UMTS and LTE Mobile Network Architectures," vol. 54, pp. 27–39, 2016.

[27] W. Paper, "Basic concepts of HSPA," *Meta*, vol. 38, no. February, 2007.

[28] C. Engineering, "Minimizing Packet Loss Using Buffer Management Scheme for Video Streaming," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 2014, no. 1, pp. 2269–2271, 2015.

[29] I. Gagro, R. Lu??a, and V. ??a??kovi??, “MPEG-4 video transfer over IEEE 802.11 WLAN,” *MIPRO 2009 - 32nd Int. Conv. Proc. Telecommun. Inf.*, vol. 2, pp. 221–225, 2009.

[30] S. M. Thampi, “A Review on P2P Video Streaming,” *ResearchGate*.

[31] S. M. Thampi, “A Review on P2P Video Streaming,” pp. 1–47, 2013.

[32] M. Rabinovich, O. Spatscheck, Q. Wang, and B. D. Davison, “Web Caching and Replication,” *SIGMOD Rec.*, vol. 32, no. 4, pp. 107–108, 2003.

[33] X. Hei, Y. Liu, and K. W. Ross, “IPTV over P2P Streaming Networks: The Mesh-Pull Approach,” *IEEE Commun. Mag.*, vol. 46, no. 2, pp. 86–92, 2008.

[34] J. G. Apostolopoulos, W. T. Tan, and S. J. Wee, “Video streaming: Concepts, algorithms, and systems,” *Handb. Video Databases Des. Appl.*, pp. 831–864, 2003.

[35] P. Antoniou, M. Lestas, A. Pitsillides and, and V. Vassiliou, “Adaptive Methods for the Transmission of Video Streams in Wireless Networks,” 2013.

[36] R. Adhikari K. and A. R.K., “An Introductory Study on Time Series Modeling and Forecasting Ratnadip Adhikari R. K. Agrawal,” *arXiv Prepr. arXiv1302.6613*, 2013.

[37] G. Bontempi and S. Ben Taieb, “Machine Learning Strategies for Time Series Forecsting,” no. January, 2013.

[38] H. D. Trinh, L. Giupponi, P. Dini, C. Cerca, A. Carl, and F. Gauss, “Mobile Traffic Prediction from Raw Data Using LSTM Networks,” no. September, 2018.

[39] A. K. Al Tamimi, R. Jain, and C. So-In, “SAM: A simplified seasonal ARIMA model for mobile video over Wireless broadband networks,” *Proc. - 10th IEEE Int. Symp. Multimedia, ISM 2008*, no. Ism, pp. 178–183, 2008.

[40] Z. H. Kilimci *et al.*, “An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain,” *Complexity*, vol. 2019, 2019.

[41] S. C. Chiemeke and A. O. Oladipupo, “Seasonal Time Series Forecasting: a Comparative Study of Arima and Ann Models,” *African J. Sci. Technol. Eng.*, vol. 5, no. 2, pp. 41–49, 2004.

[42] H. Shimodaira, “Multi-Layer Neural Networks,” *Informatics 2B*, pp. 2–4, 2015.

[43] A. P. Khadilkar, “Deep Learning Based Stress Prediction For Bottom -Up Stereo-Lithography (SLA) 3D Preinting Process,",” vol. 2, no. September, pp. 6–11, 2018.

[44] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, “Convolutional LSTM network: A machine learning approach for precipitation nowcasting,” *Adv. Neural Inf. Process. Syst.*, vol. 2015-Janua, pp. 802–810, 2015.

ANNEX

AIC, BIC and AICc values of candidate SARIMA models

Model	AIC	AICc	BIC	Remark
ARIMA(0,1,1)(0,0,2)(7)	21.27146	21.27672	20.31072	single
ARIMA(0,1,1)(1,0,1)(7)	21.23431	21.23956	20.27356	single
ARIMA(0,1,1)(1,1,1)(7)	21.25484	21.26009	20.29409	single
ARIMA(1,1,1)(1,0,1)(7)	21.23865	21.24405	20.28772	single
ARIMA(1,1,1)(1,1,1)(7)	21.25484	21.26009	20.29409	single
ARIMA(2,1,1)(1,0,1)(7)	21.24027	21.24585	20.29916	single
ARIMA(2,1,1)(1,1,1)(7)	21.25389	21.25929	20.30296	single
ARIMA(3,1,1)(1,0,1)(7)	21.24452	21.25029	20.31322	single
ARIMA(3,1,1)(1,1,1)(7)	21.25808	21.26366	20.31697	single