SCHOOL OF POSTGRADUATE STUDIES

Department of Computer Science

Ethiopian Toll Road Enterprise Traffic Flow Forecasting

By: Silashi Alamayehu Ababe

Addis Ababa, Ethiopia

July 30, 2022

SCHOOL OF POSTGRADUATE STUDIES

Department of Computer Science

Ethiopian Toll Road Enterprise Traffic Flow Forecasting

A Thesis Submitted to the School of Postgraduate Studies Presented in Partial Fulfilment of the Requirements for the Degree of Master of Science in Computer Science

By: Silashi Alemayehu Ababe

Advisor: Hailay Beyene (Ph.D.)

Addis Ababa, Ethiopia

July 30, 2022

Declaration

I, ***Silashi Alamayehu***, the undersigned, declare that this thesis entitled: ***ETRE Traffic flow Forecasting*** is my original work. I have undertaken the research work independently with the guidance and support of the research Advisor. This study has not been submitted for any degree or diploma program in this or any other institution and the sources of materials used for the thesis have been duly acknowledged.

Name of Student	Signature	Date

This is to certify that the thesis entitled: ETRE Traffic Flow Forecasting submitted in partial fulfillment of the requirements for the degree of Masters of Computer Science of the Postgraduate Studies, Admas University, and is a record of original research carried out by ***Silashi Alamayehu***, PGMGC/8018/20, under my supervision, and no part of the thesis has been submitted for any other degree or diploma. The assistance and help received during this investigation have been duly acknowledged. Therefore, I recommend it to be accepted as fulfilling the thesis requirements.

_____ _____ _____

Name of Advisor Signature Date

Certificate of Approval

This is to certify that the thesis prepared by **Silashi Alamayehu**, entitled **Ethiopian Toll Road Enterpsise Traffic Flow Forecasting,** and submitted in partial fulfillment of the requirements for the Degree of Masters of Science in Computer Science/MSc complies with the regulations of the University and meets the accepted standards concerning originality and quality.

Signature of Board of Examiner`s:

_____	_____	_____
External examiner	Signature	Date
_____	_____	_____
Internal examiner	Signature	Date
_____	_____	_____
Dean, SGS	Signature	Date

Acknowledgments

I would like to thank my advisor Hailay Beyene (Ph.D.) for his genuine pieces of advice throughout the thesis work. His suggestions are valuable and helped me to shape my work.
I would like to thank Ethiopian Toll Road Enterprise staff members, especially Ato.Mustefa Abasimel, Ato Abiy Woretaw, Ato. Tagel Ayalewu was very cooperative in providing the necessary data for this work and was very encouraging throughout the whole process.
Last but not least I also want to give special thanks to my mother Wudinesh Abera and my beloved sister Tangut Melaku and to all my parents too.
Finally, I would like to thank my friends, Efrem, and Kasaye Tamiru for giving me support, strength, and continuous encouragement throughout my work and through the process of researching and writing the thesis. This achievement would not have been possible without you. Thanks to all my family and my friends.

Abbreviations

ACF	Autocorrelation Function
ADF	Augmented Dickey-Fuller Test
AIC	Akaike information criterion
ANN	Artificial Neural Networks
AR	Autoregressive model
ARMA	Autoregressive Moving Average
ARIMA	Autoregressive Integrated Moving Average
ETRE	Ethiopian Toll Road Enterprise
DL	Deep learning
TR	Toll road
LSTM	Long-Short Term Memory
MA	Moving Average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
PACF	Partial Autocorrelation Function
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
SARIMA	Seasonal Auto-Regressive Integrated Moving Average
SM	Statistical Model

Contents

List of Tables

List of figures

Abstract

The operational planning of a toll road service depends on accurate forecasting of traffic volumes. By accurately predicting the volume of traffic at exit toll stations, staffing and scheduling levels can be set, service requirements may be met, and customer satisfaction can be improved. The Ethiopian Toll Road Enterprise currently uses the average method to predict traffic volume. However, the average approach to predicting traffic volume at an exit toll station on a toll road has difficulty since it cannot handle the trend and seasonality fluctuation. Using historical traffic flow data from Ethiopian Toll Road Enterprise, this study aims to develop a model that predicts traffic volume at exit toll stations.

For predicting and forecasting traffic volume at the exit toll stations, the researcher in this thesis suggested a time series forecasting model. SARIMA and LSTM were two of the univariate time series approaches used by the researcher. Seven years of traffic flow data is collected from Ethiopian Toll Road Enterprise. The experimental data also shows that the LSTM model has a a better RMSE and MAE improvement in forecasting error over the SARIMA model. The overall findings of this study show that the LSTM model is a useful tool for forecasting traffic volume at exit toll stations in order to represent temporal patterns. The optimization of planning, scheduling, and staffing requires such accuracy in resource allocation for toll roads.

Keywords – *Traffic flow/volume, time series, forecasting, Staffing, scheduling/planning, statistical model, deep neural network, LSTM, SARIMA*

CHAPTER ONE: INTRODUCTION

The Addis Ababa-Adama road is the main export and import trade route between Addis Ababa-Awash-Mile-Galafi (Djibouti) and Addis-Mojo-Awassa-Moyale (Mombasa) which connects the southern part of the country and Addis-Awash-Harar. As part of the Jijiga-Togochale (Berbera) corridors, it has the highest daily traffic on all of our country's highways [ERA,2014]. Due to the rapid growth of traffic, it was necessary to build an alternative road before it could negatively impact the country's economy. Accordingly, the road, which has been under construction since March 2010, has been completed and started to provide traffic flow service [(ETRE, 2022)].

As the vehicles began to use the road, alleviating traffic congestion, reducing travel time from Tulu Dimtu Adama to an average of 2 hours and a quarter to 45 minutes, significantly reducing travel time and transportation costs [ETRE, p. 2014].

The expressway is designed to be used by paid drivers. This means that the road user has to pay a portion of the maximum travel time and the cost of moving the vehicle while they were start using the highway. even though the road user needs to save time and fuel when using the toll road, this study is presented as it is necessary to provide new technology for reducing traffic congestion around the toll station for road users using the highway/toll road.

The data generated from Traffic flow have a significant effect that should be analyzed and used for decision-making purposes to determine appropriate staff levels, shift schedules, and real-time routing design for future demands [E. R. Pasupathy, 2013].

To effectively run a toll road, toll road managers must match traffic toll collection system resources to workload. The first and most important step in forecasting workload effectively is to provide an accurate forecast of future traffic flow volumes. A toll road manager frequently requires two types of projections for staffing, scheduling, and technology migration [H. Shen and J. Z. Huang, January 2003, 2014].

1. Forecast traffic volumes several days or weeks, months in advance;

2. Dynamically update the forecast on a given day based on newly available data as new traffic volume increases throughout the day.

1.1. **Background of the study**

New technologies in information and communications have had a substantial influence on our daily lifestyle and transportation is no exception. These technologies have given rise to the prospect of toll roads (TR) technology (i.e. intelligent transportation system) which aims to reduce crashes, energy consumption, pollution, and congestion while at the same time increasing transport accessibility.

Ethiopian Toll Road Enterprise (ETRE) is a publicly owned company established under the regulation NO843/2014 issued by the council of ministers in July 2014 with a paid capital of about 202 million ETB [federal negarith gazette of the FDRE,2014]. The enterprise is responsible to provide toll road services for toll road users by operating and maintaining toll roads. ETRE now manages three branches Addis-Adama with a total length of 78 km, Modjo-zuway with a total length of 92 km, and Diredawa-Dewelle with a total length of 220 km [ETRE,2022]. Those branches are designed and constructed in a new alignment and located in the country's import-export corridor. Even so, Traffic flow in ETRE was increasing periodically. During the opening of ETRE in 2014, the Addis Adama traffic flow was around 6778.6 per day [ETRE,2014]. But in 2021 the traffic volume increased to 26181.3 per day [ETRE,2021].

To operate and manage the future traffic flows in each branch, traffic flow forecasting is an important issue to decrease traffic flow congestion and find a new technology with a new diversion. This research focuses on ETRE traffic Flow forecasting in the case of Addis Adama branches [ETRE,2022].

1.2. **Statement of the problem**

Ethiopian Toll Road Enterprise provides toll road service to its customers through the semi-manual toll collection system. The semi-manual toll system is one of the toll collection systems, which delivers toll services via a tolling system and collects money manually. The services are available up to 15min to 20 min at the exit based on the best effort of manpower. It is known that semi-manual toll collection system services are used as a last-mile solution due to their delay and high service cost. However, ETRE has about 26 thousand customers such as government and private companies. The service demand and traffic volume are increasing due to the introduction of high traffic flow services [ETRE,2022].

ETRE has average method for planning traffic flow [ETRE,2022]. This makes the planning inaccurate throughout the fiscal year of strategic planning. Average method cannot handle the

trend and seasonality fluctuation. Ethiopian Toll Road Enterprise plans for traffic flow to give service using semi-manual toll collection and check for traffic congestion by forecasting the next traffic flow concerning existing traffic flow. There are challenges in awareness, reducing traffic flow congestion, adopting new technology, and forecasting traffic flow using different models and resources in ETRE.

ETRE Addis Adama Branches, now a day provides toll road service for more than 26,000[ETRE,2022] vehicles per day. According to ETRE five years, strategy plan which this research is going to find the best model to forecast, designing superefficient navigation and a safer travel system is becoming a major challenge for toll road authorities. The increase in the number of vehicles has led to an increase in traffic congestion and a decrease in travel time for travelers.

In ETRE Road Congestion is caused by multiple factors including bottlenecks manual toll collection, and road user knowledge gap. Better insights into the causes of road congestion, and its management, are vital significance avoiding minimizing congestion and providing efficient toll road service.

Road traffic flow/volume modeling, analysis, and prediction methods have been developed to understand the causes of road traffic congestion and to prevent and manage toll road traffic congestion. The forecasting or prediction of road traffic characteristics, such as speed, flow, and occupancy, allows planning of new road networks, modifications to existing road networks, or the development of new traffic control strategies.

To face this challenge:

- Accurate traffic volume prediction or forecasting is a must since it would guide the individuals to choose the best time to travel to avoid unnecessary delays due to traffic congestion or take alternative routes.
- Help ETRE to monitor the flow of traffic and make necessary arrangements to facilitate the new traffic system in that area.

1.3. **Research Question**

a. Which model is best for ETRE traffic flow forecasting?
b. Does ETRE traffic flow data is stationary?
c. What will be the proposed model for traffic flow forecasting?
d. Does the model solve the ETRE traffic flow planning problem?

1.4.**Objective**

1.4.1. General objective

The major objective of this research is to build a model for Ethiopian Toll Road Enterprise (ETRE) traffic flow/volume using univariate time series forecasting.

1.4.2. Specific objective

The specific aims of the study are:

- Understanding the operation and traffic flow/volumes of the ETRE traffic flows.
- Studying time series forecasting models for traffic flow datasets.
- selecting the appropriate model for analyzing, preprocessing, and modeling the collected time series dataset.
- Training models, including statistical and deep learning models, for collected traffic flow data.
- Testing models and assess their performance using accuracy metrics.
- Analysing performance of the models using performance matrics

1.5.**Significance of the study**

The outcome of this study has significant implications for Ethiopian Toll Road Enterprise and similar research areas. The following are the research's contributions:

- Create awareness by investigating the most commonly used time series forecasting models both statistical and machine learning in more scientific ways, and as a useful approach to forecast downloaded traffic volume and give a contribution in the area.
- Suggest the most accurate model that improves the performance of forecasting to be used as input for ETRE in traffic flow planning and optimization.
- The research findings can be used as a reference or benchmark for future related works.

1.6. **Scope of the study**

In this study, we are going to cover the ETRE traffic data analysis, modeling of time-series dataset, and identify the best forecasting model for future Traffic flow planning and optimization tasks to ETRE.

1.7. **Limitations of the study**

The research is restricted to traffic volume data, not covering the video Data. Besides, the collected data is limited to ETRE Addis Ababa - Adama Branches. The researcher does include the other ETRE branches due to a lack of complete data.

1.8. **Definition of Terms**

Ethiopian Toll Road Enterprise: - the name of the governmental profitable company which governs all the toll roads in Ethiopia [ETRE,2014].

Toll Road: A toll road, also known as a tollway, or as a turnpike in the United States, is a public or private road (almost always a controlled-access highway), for the use of which a fee (or toll) is paid. It is a form of road pricing, typically implemented to help recoup the costs of road construction and maintenance (wikipedia, n.d.).

Tolling: remains the preferred way of funding expensive transportation infrastructure investments that are vital to support global economic growth. In addition, it helps fund the development of newer and more sustainable modes of transport. With a well-designed, fit-for-purpose tolling strategy, transportation agencies can both support their development plans while providing a better experience for road users (KapschTrafficCom, n.d.)

Traffic flow: Flow is one of the most common traffic parameters. Flow is the rate at which vehicles pass a given point on the roadway, and is normally given in terms of vehicles per hour. The 15-minute volume can be converted to a flow by multiplying the volume by four. (ITS, 2022)].

Traffic flow forecasting: means forecasting the volume and density of traffic flow,

Volume: -Volume is simply the number of vehicles that pass a given point on the roadway in a specified period (Traffic Flow Theory, 2022).

Forecasting: - is the process of making predictions based on past and present data.

ITS: - Intelligent Transportation System (ITS) can be generally defined as the application of electronics, communications, and information technology used as an integrated traffic management system to improve the efficiency or safety of the surface transportation system.

Variable Message Sign (VMS): - is a sign capable of displaying pre-defined or freely programmable messages which can be changed remotely with individual pixel control. to

provide information and advice to drivers and riders independent of any in-vehicle systems. In ETRE VMS signs can replace fixed roadside signs to inform drivers of speed limits – if necessary with the flexibility of changing mandatory speed limits, and safalertslert in response to traffic or road conditions where the legislation allows (ITS, 2022).

Tolling system: - is the system that ensures that the most economical and effective Demand Management solutions are deployed and operated to manage continued traffic growth.

Toll station: - is the location or place at which customers enter or exit toll roads.

Traffic management center (TMC): - serve as the mission control for a toll road area's major street and highway network. This one location monitors traffic signals, intersections, and roads and proactively deploys traffic management strategies to reduce congestion and coordinate state and local authorities during special events, emergencies, or daily stop-and-go traffic [ETRE,2014].

1.9. **Organization of the research**

This research is organized into five chapters.

Chapter one: is an introductory chapter describing the background, statement of the problem, objective, hypothesis of the study, scope of study, limitation of the study, significance of the study, definition of terms as well organization of the study.

Chapter two: literature review deals with **t**he review of the research of various studies carried out in the toll road and other related areas with an emphasis on forecasting practices.

Chapter three: explains the research design, methodology, techniques of data collection, method of analysis, and presentation.

Chapter four: deals with the proposed model for Ethiopian Toll Enterprise Traffic flow forecasting.

Chapter Five: deals with Results and Discussion.

chapter six: conclusion and Recommendations. This chapter presents a summary of the whole study with inferences and discusses the summary, conclusion, and recommendation. The scope for future research will also be discussed.

CHAPTER TWO: LITERATURE REVIEW

2.1. Theoretical Concept of Traffic flow forecasting

To manage organizations effectively and efficiently, every function of traffic flow forecasting shall be given due consideration. Particularly those functions of the organization taking a significant portion of the traffic congestion rather require preferential attention. Traffic Safety and giving qualified toll service is one of the basic functions of all types of toll roads. It is basically because no toll road can operate without it. Thus, the success of any toll road depends on decreasing traffic flow congestion and providing qualified toll road service by forecasting traffic volume as it does on the executives who administer the other function of the organization. For this reason, forecasting ETRE traffic flow/volume by using secondary data and reviewing related thesis and journals is important to select and identify the best model for ETRE toll road service planning.

Forecasting road traffic flow conditions is essential for advanced traffic management information systems, which mainly aim to reduce traffic congestion and improve mobility of transportation. Short-term traffic flow forecasting, which has a horizon of only a few minutes, is highly suitable for traffic management information systems in supporting proactive dynamic traffic control to anticipate traffic congestion [V. B. Arem, H. R. (Mar. 1997], [S. Innamaa, Apr. 2006], [A. Simroth and H. Zähle, Mar. 2011]. Long-term forecasting can be used for planning and technology migration across the toll roads. Getting an accurate prediction of the future states and conditions of traffic is an attractive topic for many researchers in the field of ITS. Having the ability to predict traffic attributes such as speed, travel time, and traffic flow, plays an important role in various components of ITS such as Advanced Traveler Information Systems (ATIS), or Advanced Traffic Management Systems (ATMs) [Vanajakshi and Rilett,2004]. In this chapter, the researcher is going to select the best model by reviewing different articles, journal research, etc...

2.2. Theory about Toll Road

Tolls have been placed on roads throughout history, typically to generate funds for the repayment of toll revenue bonds used to finance construction or operation [wikipedia, n.d.]. As previously stated, Ethiopia has the first toll road from Addis Abeba to Adama, and most road users refer to this toll road as an expressway/highway. Customers chose this toll road to save time and fuel, even though they also chose it for its safety and service quality [ETRE,2013]. As a result, the toll road must be safe, technologically efficient, and dependable in order to meet the needs of customers promptly [ERA,2013]. There are different types of toll roads:

a. An open toll road is one with no guardrail between the lanes and no compound on both sides of the road. The customer pays at the entrance toll station and can leave without having to pass through the next toll station. This toll road is illustrated by the Diredawa-Dewelle branch [ETRE,2019].

b. A Closed toll road is a closed toll road with compound on both sides of the road in both directions and possibly guardrails between toll lanes. There are entrance lanes on this type of toll road. Customers purchase a ticket at the toll station's entrance and pay at the exit toll station. Toll roads of this type include the Addis-Adama and Modjo-Batu branches [ETRE,2014].

2.3. Studying Addis-Adama Toll Road Branch

2.3.1. Addis -Adama Toll Stations

ETRE in the case of Addis-Adama branch has six toll stations. Each toll station has entrance and exit to provide toll road service. The data collected at each toll station can be stored in a central database at Tulu Dimtu's head office [ETRE,2014]. Figure 2 shows the general LAN architecture of ETRE [Appendix 5 and 6]. Figure 1 show the name of each toll station and their kilometers within toll lane number at the exit and entry of toll booth at each toll station.

Figure 1 : Name of Addis Adama Toll Stations

S.no.	Station name	Entry no	Exit
3.	Tulu-Dimtu Toll Station (K2)	4	9
4.	Bishoftu North(k16)	2	3
5.	Bishoftu South(k33)	2	3
6.	Modjo(k52)	3	3
7.	Adama west A(k60)	2	2
8.	Adama west B(k60)	2	2
9.	Adama main(K64)	4	7

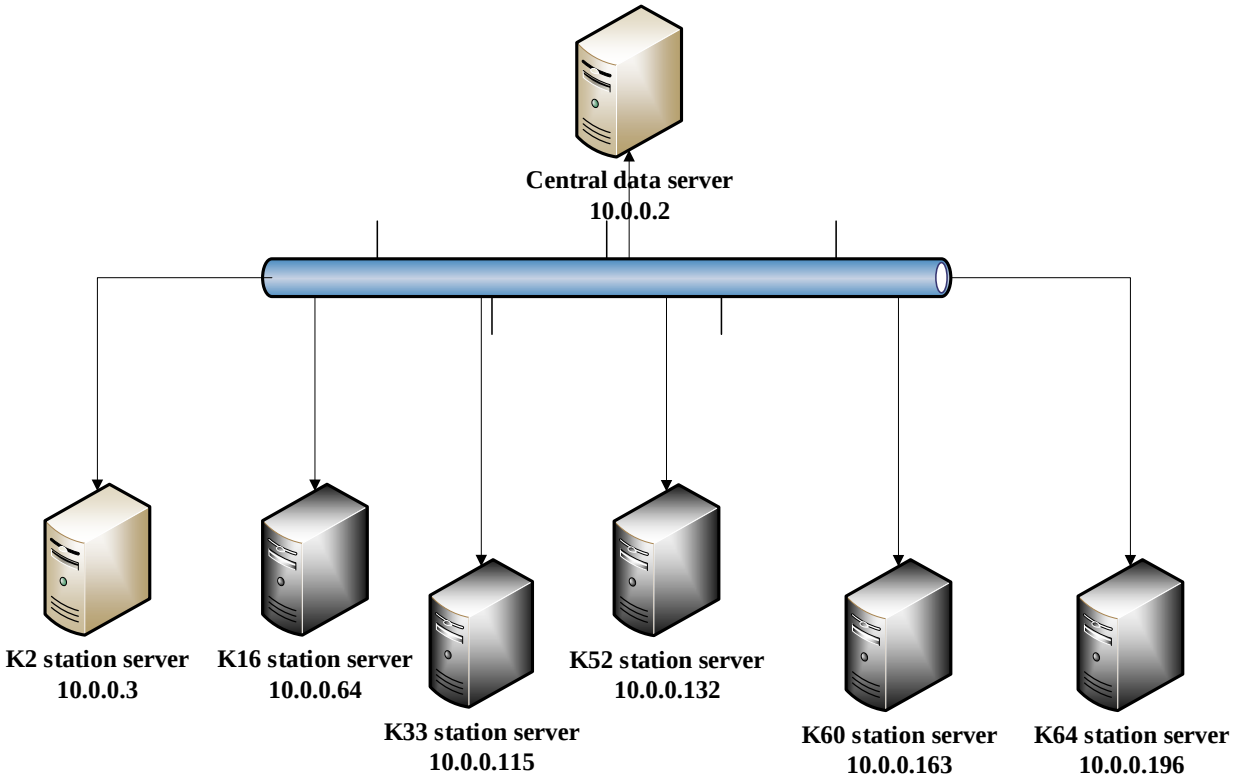


Figure 2: ETRE LAN ARCHITECTURE

As previously stated, traffic flow data was collected at each toll station and stored with the respective station server before being forwarded to the central data server for use [ETRE,2014]. Then the data that was sored in station server can be sent central server [Appendex no. 5 and 6]. This researcher focuses on exit traffic flow data rather than entry traffic flow data because ETRE uses exit data for planning and traffic congestion can be seen at the exit rather than the entrance in Addis-Adama

2.3.2. Toll System Business Structure

The Toll System Business Structure includes toll center, toll branch center, toll station, and toll lanes, and the corresponding operation and management organization are ETRE, AAE sub-center, and toll station (toll lanes), as shown in the following figure 3:

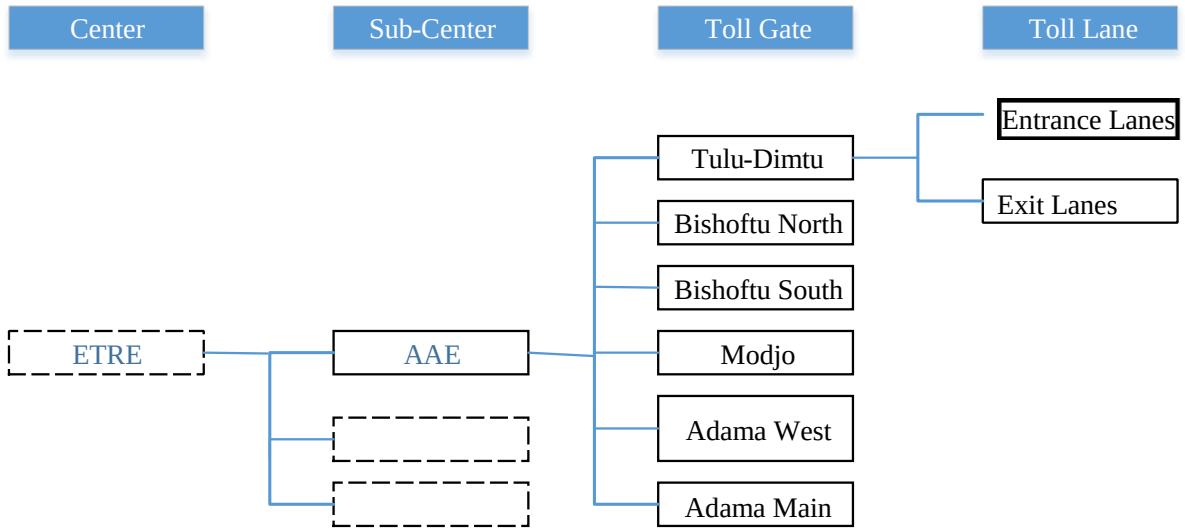


Figure 3 ETRE Toll system Business Structure

2.4. Time series forecasting

Time series forecasting is the process of analyzing time series data using statistics and modeling to make predictions and inform decision-making strategies [(Zhao, Chen, Wu, Chen, & Liu, 2017)]. It is not always accurate, and the probability of forecasting can vary dramatically — especially in the face of frequent fluctuations in time series data and factors beyond the control. However, the understanding of prediction about which results are most likely — or less likely to occur than other possible consequences. Often, the more the researcher uunderstands the more predictions can be made. Although forecast and “forecasting” often mean the same thing, there are significant differences. In some industries, forecasting may refer to data at some point in the future, whereas forecasting refers to future data. Series predictions are often used in conjunction with time series analyses. Time series analysis involves developing models to gain data comprehension to understand the underlying causes. Analysis can provide "why" after the results are seen. Predicting then takes the next step in what to do with that information and predictable interpretation of what might happen in the future.

2.4.1. Definition of time series

A time series is a sequence of data points in which the orders indicate the sequence of measurable values over time. A time series is defined mathematically by the values x1, x2, x3…...xn of the variable x at times t1, t2, t3...tn. A time series may be continuous or discrete. A continuous time series includes observations taken at every point in time, whereas a discrete time series includes observations taken at discrete points in time. A univariate time series is

one that only contains observation data for a single variable. However, when records from more than one variable are considered, it is referred to as multivariate time series [(Agrawal)].

2.4.2. Time series analysis

Time series analysis is the process of analyzing time series data to extract meaningful statistics and other data characteristics. Time series are studied for a variety of reasons, including forecasting the future based on past information, understanding the phenomenon underlying the measures, or simply providing a concise description of the series' significant features [(H. Zou and Y. Yang, 2004)].

2.4.3. Components of Time Series

Time series data contains a variety of patterns, and it is often helpful to split a time series into several components, each representing a different underlying pattern category. Often this is done to help improve understanding of the time series, but it can also be used to improve forecast accuracy. A time series can be decomposed into four major components: trend, seasonal, cyclical, and residual; each of them is described as follows [(Davis)].

2.4.4. Trend Components

The trend is a non-repeating and long-term pattern of a time series. A trend can be either positive or negative depending on whether the time series shows an upward long-term pattern or a downward long-term pattern.

2.4.5. Seasonality Components

Seasonality describes a repeating behavior that occurs over the short term at a predictable interval, which spans less than a year, such as hourly, weekly, monthly, or quarter.

2.4.6. Residual Components

The residual component is unpredictable. Every time series has some unpredictable component 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.

2.4.7. Cyclic Components

A cyclical pattern is defined as any pattern that shows an up and down movement around a given trend. The length of a cycle is determined by the type of business or industry being studied.

For decomposed time series data, two types of mathematical models are typically used to account for the effects of trend, seasonality, residual and cyclic components. These models are

additive and multiplicative. The additive model works based on the assumption that the four components are independent of one another, whereas the multiplicative model works based on the assumption that the four components of a time series are dependent on one another. Mathematically represerepresentations additive and the multiplicative models are described in Equations (2.1) and (2.2) respectively.

Equation 1 Mathematical Representation of the additive

$$Y(t) = T(t) + S(t) + C(t) + R(t) \quad (2.1)$$

Equation 2 Multiplicative model equation

$$Y(t) = T(t) \times S(t) \times C(t) \times R(t) \quad (2.2)$$

Where Y(t) is the observation and T(t) + S(t) + C(t) + R(t) are the trend, seasonal, cyclical, and residual at time t respectively.

2.5. Time Series Forecasting

A time series forecasting method is a technique for predicting the future value based on current and historical data. In general time series forecasting techniques are classified into a linear model, nonlinear model, hybrid model, and decomposed model (H. Zou and Y. Yang, 2004)].

- **Linear time series models:** linear time series model is to study the dynamic structure of such a series of data. The two main subgroups of this technique are Auto-Regressiveive (AR) and Moving Average (MA) models. The autoregressive moving average model (ARMA) and seasonal autoregressive integrated moving average (SARIMA) model are created by combining these two models.
- **Nonlinear time series models:** are used to investigate aspects that linear processes can't handle, such as cycle and time-change variance. Techniques such as Neural Networks are one example [(I. Khandelwal R. A., 2015)].
- **Hybrid model:** It is mostly made up of a mix of linear and nonlinear models. For instance, ARIMA with an Artificial Neural Network.
- **Decomposed model:** is implemented by splitting down the time series into seasonal, trend, cyclical, and irregular components. For example, the nonlinear decomposed model decomposes time series into trend, period, mutation, and random components.

2.5.1. Statistical Forecasting Model

A statistical model is a mathematical representation of observation data that attempts to analytically determine the relationship between two or more random variables. Forecasting time series models are either 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 time series data models can take many forms and represent various stochastic processes [(H. Zou and Y. Yang, 2004)]. Some examples of statistical forecasting models include the autoregressive (AR), moving average (MA), and ARIMA models. Box and Jenkins [(B. De Constance and B. De Constance)] proposed a very successful variation of the ARIMA model for seasonal time series forecasting, namely the Seasonal ARIMA (SARIMA) model [(P. J. Brockwell and R. A. Davis)][(B. Sennaroglu and G. Polat, no. JUL, vol. 2017,)][(M. Milenković, no. April, 2016,)][(P. P. Dabral and M. Z. Murry, 2017,)].

In this section, the researcher discusses some of the linear statistical models with mathematical expressions.

2.5.1.1. AR Model

AR model forecasts future variables based on past values. AR models are based on the idea that the current value of the series, Xt, can be explained as a linear combination of p past values, Xt-1 +Xt-2+...Xt-p +Wt where Wt is White noise in the same series. AR (p) is a common notation for an AR model, where p is the model order. The term "auto-regression" refers to a regression of the variable itself. The AR model's order indicates how many lagged past values are included. The mathematical description is shown in Equation (2.3).

$$\mathrm{x(t)} = \sum_{\mathrm{j=1}}^{\mathrm{P}} a_j x(t-j) + w(t) \qquad (2.3)$$

Where,

- $x\,(t-j)$ is the previous values sample;
- a_j is autoregressive coefficients at the order of p; $w(t)$ is White noise with zero mean.

2.5.1.2. Moving Average Model

The MA model, rather than using past forecast values in a regression, uses past forecast errors in a regression-like model. The MA model considers a sample process to be a moving average (unequally weighted) of a random sample process. The moving average model's order is denoted by the letter q. The common notation of MA is MA(q). The mathematical description is shown in Equation (2.4).

$$x(t) = \sum_{j=1}^{q} b_j w(t-j) + w(t) \tag{2.4}$$

Where,

$w\ (t\ -\ j)$ is the previous values sample; b_j is autoregressive coefficients at the order of p; $w(t)$ is white noise with zero mean.

2.5.1.3. ARMA Model

ARMA model is obtained from two models (a hybrid of an autoregressive and a moving average model). In time series analysis, the ARMA model is used to describe stationary time series. The ARMA model represents time series produced by sequentially passing white noise through a recursive and a non-recursive linear filter. The common nation of the ARMA model is ARMA (p, q); where:

p is the order of the autoregressive polynomial,

and q is the order of the moving average polynomial.

The ARMA model of mathematical description is shown in Equation (2.5).

$$x(t) = w(t) + \sum_{j=1}^{q} a_j x(t-j) + \sum_{j=1}^{q} b_j w(t-j) \tag{2.5}$$

2.5.1.4. ARIMA Model

The ARIMA model is one of the most widely used and well-known stochastic time series models. The integrated ARMA (ARIMA) is a type of ARMA that includes differencing [(H. Zou and Y. Yang, "Combining time series models for forecasting," ,doi: 10.1016/S0169-2070(03)00004-9. , 2004)]. ARIMA model is the most general class of models for forecasting a time series which can be made to be "stationary" by differencing (if necessary), perhaps in conjunction with nonlinear transformations such as logging or deflating (if necessary). The

ARIMA model can be viewed as a “filter” that tries to separate the signal from the noise, and the signal is then extrapolated into the future to obtain forecasts. The ARMA procedure offers a comprehensive set of tools for univariate time series model identification, parameter estimation, and forecasting, as well as a high level of flexibility in the types of ARIMA models that can be analyzed. Seasonal, subset, and factored ARIMA models are supported, as are intervention or interrupted time series models, multiple regression analysis with ARMA errors, and rational transfer function models of any complexity [(B. De Constance and B. De Constance)].

The ARIMA model is an ARMA model that is based on a differenced series, which is created by combining differencing with AR and MA models. The ARIMA model is denoted as the ARIMA (p, d, q) model, where p represents the number of autoregressive terms, d represents the number of non-seasonal differences required for stationary, and q represents the number of moving average terms. The mathematical description of the ARIMA model is shown in Equation (2.6).

$$x(t) = c + \sum_{j=1}^{q} a_j x(t-j) + \sum_{j=1}^{q} b_j w(t-j) + w(t) \qquad (2.6)$$

2.5.1.5. SARIMA Model

SARIMA models are an extension of ARIMA that explicitly supports univariate time series data with seasonal components. This model is divided into two parts: non-seasonal and seasonal, with the AR, Integrated, and MA parameters in each. It is used when there is a periodic characteristic in the data that must be known ahead of time. SARIMA (p, d, q) (P, D, Q) s is the general form of a seasonal ARIMA model, where p is the non-seasonal AR order, d is the non-seasonal differencing, q is the non-seasonal MA order, P is the seasonal AR order, D is the seasonal differencing, Q is the seasonal MA order, and S is the time steps of repeating seasonal pattern.

1. SARIMA modeling steps

The most crucial step in estimating the SARIMA model is determining the values of seven parameters (p, d, q) (P, D, Q, s). If the variance grows with time, for example, based on the time plot of the data, we should use variance-stabilizing transformations and differences. Then, using the ACF to determine how much linear dependence exists between observations in a time series separated by a lag p and the PACF to determine how many autoregressive terms q is

required, we can identify the preliminary values of autoregressive order p, order of differencing d, moving average order q, and their corresponding seasonal parameters P, D, and Q.

a. Stationary process

Many time series techniques assume that the data are stationary. A stationary process has the statistical property that the mean, variance, and autocorrelation structure do not change over time [(P. J. Brockwell and R. A. Davis)]. The requirements for achieving stationarity necessitate the fulfillment of the three constraints listed below:

- The mean value has to be approximated to a constant: $E(X_t)$= constant for all t;
- The variance has to be approximated to a constant: Variance (X_t)= constant for all t;
- The co-variance must be dependent on lag j: Cov $(X_{t,} T_{t+j})$ =constant for all t.

b. Augmented Dickey-Fuller Test

The Augmented Dickey-Fuller test (ADF test) is a unit root test that is used to determine whether or not a given time series is stationary. It is one of the most commonly used statistical tests for analyzing the stationary of a series. To perform an Augmented Dickey-Fuller test in python it returns the p-value.

The null hypothesis of the ADF test is that the time series is non-stationary. So, if the p-value of the test is less than the significance level (0.05) then reject the null hypothesis and infer that the time series is indeed stationary. So, in this thesis, if P Value > 0.05, the researcher go ahead with finding the order of differencing. When the P-Value > 0.05, the null hypothesis is rejected and the time series is assumed to be stationary. The null hypothesis assumes that the time series is non-stationary.

c. Autocorrelation Function and Partial Autocorrelation Function plot

The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots graphically summarize the strength of a relationship between an observation in a time series and observations at previous time steps. To investigate the linear relationship between two variables, the ACF is used. The ACF is a method for calculating the linear relationship between an observation at time t (the current time) and previous observations. The PACF summarizes the relationship between an observation in a time series and observations at previous time steps by removing the relationships of intervening observations.

d. Akaike Information Criterion

The Akaike Information Criterion (AIC) is a mathematical formula for assessing the accuracy of a model. AIC is used in calculations to compare different possible models and determine which one best fits the data. AIC is calculated using the number of predictive variables used to model a model and the maximum probability of a model (the way the model generates data).

2.5.2. Deep Learning Forecasting Model

Deep learning (DL) algorithms are a more advanced and mathematically complex evolution of machine learning algorithms. DL is a subfield of machine learning, with neural networks serving as the foundation of deep learning algorithms. It analyzes data in the same way that humans do. DL methods, such as automatic learning of temporal dependence and automatic handling of temporal structures such as trends and seasonality, hold a lot of promise for time series forecasting. DL models for time series forecasting come in a variety of flavors such as the Artificial Neural Network (ANN) model, Recurrent Neural Network (RNN) model an LSTM model.

2.5.2.1. Artificial Neural Network Model

The ANN model is a type of intelligent system that can be used to solve complex problems in a variety of applications, such as optimization, prediction, modeling, clustering, pattern recognition, simulation, and more [(U. Yolcu, pp. 1340–1347, 2013)][(I. Khandelwal R. A., “Time series forecasting using hybrid arima and ann models based on DWT Decomposition,”, doi: 10.1016/j.procs.2015.04.167. , 2015,)]. Three layers of artificial neurons or nodes make up the ANN structure: an input layer that collects data, an output layer that computes information, and one or more hidden layers that connect the input and output layers. Each of the ANN layers contains neurons, which are nodes. A layer is a collection of nodes that have the same input and output connections but don't communicate with one another in the same layer. Figure 4 shows the ANN general architecture with three layers.

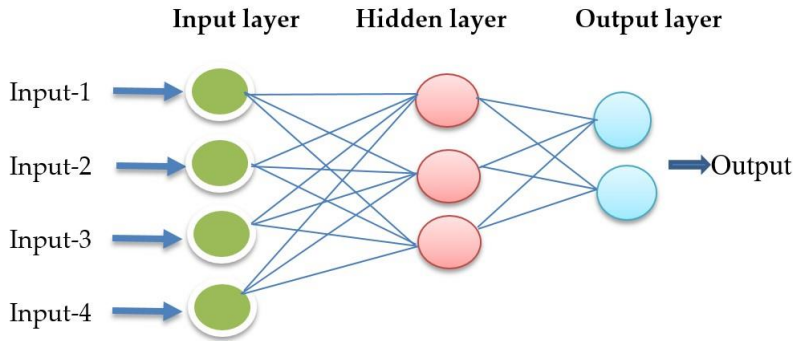


Figure 4: ANN architecture with 3 layers [*(I. Khandelwal R. A., 2015,)*]

A neuron is a mathematical function that simulates the operation of a biological neuron within an artificial neural network. A neuron typically computes the weighted average of its input and then passes this sum through an activation function. The number of features in the dataset determines the dimensionality, or the number of nodes, in the input layer. "Synapses" connect these nodes to the nodes created in the hidden layers. The synapse links have some weights for each node in the input layer. The weights function as a decision maker, determining which signals or inputs may or may not pass through. The weights also represent the strength and depth of the hidden layer. A neural network learns by varying the weight for each synopsis.

2.5.2.2. Recurrent Neural Network Model

RNN is a class of artificial neural networks [(J. Bayer and C. Osendorfer, 2014,)] that is commonly used in natural language processing and time series forecasting. The RNN model employs sequential observations, to predict the next step in the sequence of observations based on the previous steps observed in the sequence. The RNN model adds a hidden state that is generated by the sequential information of a time series, with the output dependent on the hidden state. Figure 5 shows the most common architecture of the RNN model begins to unfold into a fully connected network.

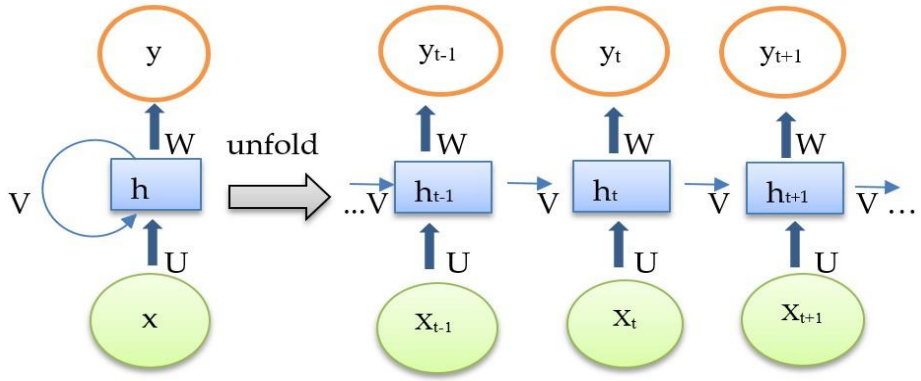


Figure 5 : RNN model architecture [(J. Bayer and C. Osendorfer, pp. 1–9, 2014)]

The mathematical expression for the RNN model in Figure 5 is denoted in Equations (2.7) and (2.8):

$$h_t = f(Ux_t + Vh_{t-1}) \quad (2.7)$$

$$y_t = f(Wh_t) \quad (2.8)$$

Where;

- X_t is the input at time t;
- U denotes the weight matrix from the input layer to the hidden layer;
- V denotes the weight of recurrent computation;
- *W* denotes weight from the hidden layer to the output layer;
- h_t denotes values of hidden nodes at the time of t;
- y_t denotes a value of the output node at the time of t;
- f is the activate function, which has many alternatives such as the sigmoid function and ReLU.

The RNN model is well-modeled time series data that is the temporal dependency [(J. Bayer and C. Osendorfer, "Learning Stochastic Recurrent Networks,", 2014)]. The main problem with a typical generic RNN is that these networks remember only a short-term dependence in the sequence. So, it is hard to remember longer dependency within sequences of data due to the vanishing gradients problem. This problem is solved by utilizing the "memory" introduced in the LSTM recurrent network [(H. Palangi, pp. 4504–4518, 2016,)].

2.5.2.2.1. LSTM Model

The LSTM model is a type of recurrent neural network that can remember patterns selectively for long periods. The hidden layer unit in the original RNN architecture is replaced by a memory block (called cells) in the LSTM architecture. It carries all the information with only some linear interaction. The LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called gates. The gates, which are based on a sigmoid neural network layer, allow cells to either pass data through or discard it. Each sigmoid layer produces a number between 0 and 1, indicating how much of each data segment should be allowed through in each cell. A value of zero means "let nothing through," while a value of one means "let everything through!" [(W. Bao J. Y., pp. 1–24, 2017,)].

LSTM networks are typically made up of memory blocks known as cells that are linked together via layers. Cell state Ct and hidden state h_t contain information, which is regulated by mechanisms known as gates via sigmoid and tanh activation functions. As a result, LSTM can conditionally add or delete information from the cell state. In general, the gates take the hidden states from the previous time step h_{t-1} and the current input x_t as inputs and multiply them pointwise by weight matrices with a bias added to the product.

The LSTM network unit consists of a cell, an input gate, an output gate, and a forget gate. A cell retains values for an autocratic time interval. The input gate controls the flow of information into the cell. The output gate controls the flow of information to and from the outside world. Similarly, forget gates control the flow of information that is necessary or unnecessary. All of the LSTM network units are provided in detail below:

A. Forget gate

It aids in determining whether information can pass through the network's layers. It expects two types of input from the network: information from previous layers and information from the presentation layer. In general, the forget gate layer decides what information from the previous cell state to retain. The mathematics is represented in Equation (2.9).

$$f_t = \sigma(w_f * [h_{t-1}, x_t] + b_f) \quad (2.9)$$

B. Input gate

The following step is to decide what new information we will store in the cell state. This is divided into two parts. First, a sigmoid layer known as the "input gate layer" determines which

values to update. A tanh layer then generates a vector of new candidate values, ct, that could be added to the state. In the following step, we'll combine these two to create a state update. The information that must be stored is then determined by the following two Equations.

$$i_t = \sigma(w_i * [h_{t-1}, x_t] + b_i) \quad (2.10)$$

$$c_t = tnah(w_c * [h_{t-1}, x_t] + b_c) \quad (2.11)$$

C. Output gate

Finally, the output gate is the last gate that helps decide how much information from the current cell state flows into the hidden state. This output will be based on our cell state but will be a filtered version. First, run a sigmoid layer which decides what parts of the cell state then go to output.

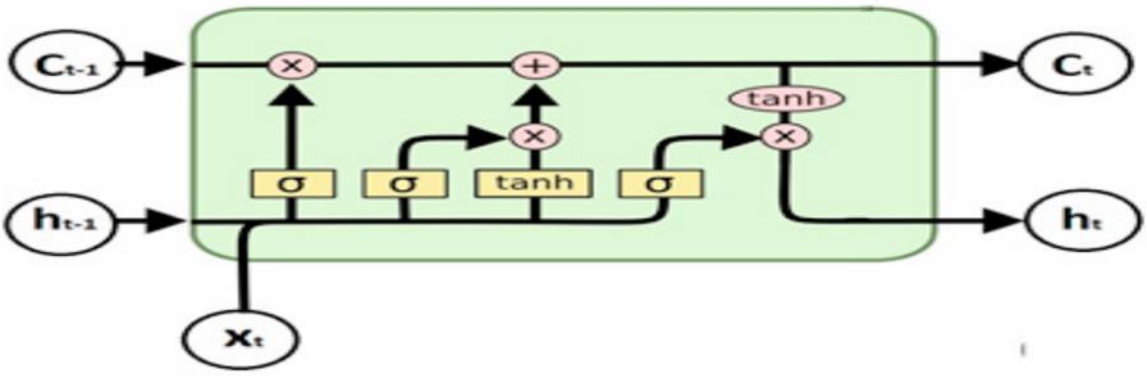


Figure 6: LSTM Architecture [*(W. Bao J. Y., 2017,)*]

Then, the researcher can put the cell state through tanh (to push the values to be between −1 and 1) and multiply it by the output of the sigmoid gate, so that output the parts that are decided. The mathematical representation in Equations (2.12) and (2.13).

$$o_t = \sigma(w_o * [h_{t-1}, x_t] + b_o) \quad (2.12)$$

$$h_t = tnah(C_t) \quad (2.13)$$

Each LSTM typically employs one of three types of gates to regulate the state of each cell: The Forget Gate generates a value between 0 and 1, with 1 denoting "totally keep this" and 0 denoting "absolutely ignore this." – Memory Gate selects the new information that has to be stored in the cell. The values that will be adjusted are first selected by an input door layer sigmoid layer. After that, a tanh layer creates a vector of potential new values to be added to

the state. – What each cell will give is determined by the output Gate. The value that is produced will be dependent on the cell status as well as the newly added and filtered data.

In the literature, several statistical models have been proposed to forecast time series for Exit volume traffic in the context of toll road service. ARIMA, Holt-Winters smoothing, and SARIMA models are the most widely used univariate forecasting methods (J. W. Taylor and J. W. Taylor, no. September 2015, 2008). The data generating function is constrained by the inherent constraint of linearity in these models. To address this, several nonlinear models have been developed in the literature.

Deep learning algorithms, in particular, have introduced new approaches to prediction problems in which the relationships between variables are modeled in a deep and layered hierarchy. Deep learning-based algorithms like LSTM which is a of RNN have gotten a lot of attention in recent years due to their applications in a variety of fields, including finance. Deep learning methods can detect data structure and pattern, such as nonlinearity and complexity in time series forecasting (W. Bao J. Y., 2017,).

The main goal of this review different literature to select the best model by comparing comparing statistical and deep neural network models for forecasting Traffic flow volumes, using the SARIMA and LSTM algorithms. Which forecasting methods provide the best predictions in terms of lower forecast errors and higher accuracy of forecasts. Using these proposed methods, the researcher does the experiment and obtains the required results in the next chapter.

2.6. Related work

Traffic flow forecasting is important for toll road planning, improving traffic efficiency, safe roads, reducing traffic congestion, toll road traffic management, reducing travel time, and increased road capacity [(P. Ross, 1982,)].

Forecasting traffic is the basis of planning and construction of transport facilities and also it accurately plays an important role in the healthy development of transportation. The physical characteristics and economic characteristics of the region should be addressed for transport infrastructure planning. Various factors affecting transport planning such as socio-economic, demographic, and travel demand requirements of the region are to be studied. The pattern of traffic growth rate, projected traffic volume and economic growth rate are the major factors in highway project analysis [Rohit Galani1,2020].

Single-point short-term traffic flow forecasting will play a key role in supporting demand forecasts needed by operational network models. Seasonal autoregressive integrated moving average (ARIMA), a classic parametric modeling approach to time series, and nonparametric regression models have been proposed as well suited for application to single-point short-term traffic flow forecasting [Brian L. Smith a, 2002]. Past research has shown seasonal ARIMA models to deliver results that are statistically superior to basic implementations of non-parametric regression. However, the advantages associated with a data-driven nonparametric forecasting approach motivate further investigation of refined nonparametric forecasting methods. Following this motivation, this research effort seeks to examine the theoretical foundation of nonparametric regression and to answer the question of whether nonparametric regression based on heuristically improved forecast generation methods approaches the single interval traffic flow prediction performance of seasonal ARIMA models [Brian L. Smith a, 2002].

The selection of a model suitable for a specific purpose depends on several factors including the context of forecasts, the availability of historic data, the degree of accuracy desirable, the cost of the evaluation, and the period to be forecast. (S. Innamaa, Apr. 2006.)

Several researchers studied statistical and machine learning time series modeling and forecasting of flow data traffic. Some of them are reviewed and presented as follows.

Traffic flow forecasting has become one of the main tasks in the field of smart transport systems [Lippi, M., Bertini, M., & Frasconi. (P.2013)]. Statistical methods, artificial intelligence, and data mining techniques have been progressively used in recent years to analyze data in road traffic and predict future traffic indicators. Previous research indicates that there is no single technology that is capable of analyzing large datasets only by itself. Therefore, depending on the data structure and its volume, it is necessary to apply the appropriate technology to get the best insight from the collected data. [Sun, Shiliang, et al.] proposed a new approach to Bayesian traffic flow forecasts. A Bayesian network is the movement of traffic between the related transport links. A Gaussian mixture model (GMM) with parameters estimated using the CEM algorithm is described as a joint probability distribution of the core (data used for projections) and the core (data expected).

A deep learning approach to forecast traffic flow for short intervals on road networks is proposed in [Zhao, Z.; Chen, W.; Wu, X.; Chen, P.C.; Liu,2017]. The authors employed a traffic flow forecast system based on long short-term memory (LSTM) for prediction. The

training algorithm takes an origin-destination correlation (ODC) matrix as input. The dataset for this method was obtained from the Beijing Traffic Management Bureau and consists of about 26 million records collected from more than 500 observation stations or sensors. Data was collected at five-minute intervals from 1 January 2015 to 30 June 2015, with the first five months' data being used for training and the rest for testing. Mean absolute error (MAE), mean square error (MSE), and mean relative error (MRE) are used to evaluate the suggested model. The flow is predicted using input data at 15-, 30-, 45-, and 60-minute time intervals. The authors chose three observation stations with high, medium, and low flow rates to compare the actual and anticipated flow values at those points. The MRE values reported in this study for 15-minute interval flow prediction are 6.41 percent, 6.05 percent, and 6.21 percent. They compared the results to other methodologies such as RNN, ARIMA, SVM, RBF, and so on, and found that RNN is pretty reliable for time intervals less than 15 minutes, but error increases with larger time intervals. It outperforms other older machine learning models in general. As a result, this researcher approved that LSTM is a good solution for long-time intervals.

Theja and Vanajakskshi (2010) investigated the application of Support Vector Machines (SVM) for the short-term prediction of traffic parameters, namely speed, space headway, and volume under heterogeneous traffic conditions. Sensitivity analysis was performed to find optimum parameters of support vector regression (SVR) in terms of accuracy and running time. A comparison of performance was carried out between SVM and the multilayer feed-forward neural network with backpropagation. Recent studies (Centiner et al.,2010; Hu et al., 2010; Pamula, 2011) applying different ANN architectures with different input parameters measured through field studies by using advanced instruments demonstrate that ANN modeling is an effective approach for short-term traffic flow modeling. A novel neural network (NN) training method that employs the hybrid exponential (EXP) smoothing method and the Levenberg Marquardt (LM) algorithm was developed by Chan et al. (2012), which aims to improve the generalization capabilities of previously used methods for training NNs for short-term traffic flow forecasting. The proposed method was evaluated by forecasting short-term traffic flow conditions on the Mitchell Freeway in Western Australia. Results indicate that, in general, test errors obtained by EXP-LM are smaller than those obtained by the other tested algorithms. It was concluded that NNs with superior generalization capabilities for traffic flow forecasting can be obtained by using EXP-LM.

Figure 7: List of some of the reviewed article for selecting models

sno	Author	Title	Used Models	Year
1	J. W. Taylor and J. W. Taylor, no. September 2015, 2008	A Comparison of Univariate Time Series Methods for Forecasting Intraday Arrivals at a Call Center	ARIMA, Holt-Winters smoothing, and SARIMA models	2008
2	W. Bao, J. Yue, and Y. Rao,	A deep learning framework for financial time series	stacked autoencoders and LSTM	2017
3	J. Bayer and C. Osendorfer,	"Learning Stochastic Recurrent Networks,"	SRN	2014
4	U. Yolcu, pp. 1340–1347, 2013	"A new linear & nonlinear arti fi cial neural network model for time series forecasting,	Artificial neural network	2013
5	S. Innamaa	Effect of monitoring system structure on short-term prediction of highway travel time	LSTM	2006
6	Zhao, Z., Chen, W., Wu, X., Chen, P., & Liu, J.	A deep learning approach for short-term traffic forecast.	LSTM network	(2017).
7	J. W. Taylor and J. W. Taylor.	A Comparison of Univariate Time Series Methods for Forecasting Intraday Arrivals at a Call Center	LSTM	2008
8	J. Brownlee. (2016).	"Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras	LSTM	2016

2.7. Summary

There are different time series forecasting methods used for various applications such as ARIMA, Holt-Winters smoothing, SARIMA, and seasonal moving average models in traditional forecasting methods [(D. K. Barrow, 2016)]. Other modern deep learning learning time series forecasting methods are multilayer perceptron (MLP) and LSTM. The statistical forecasting methods handle only linear data behavior. Whereas, deep learning algorithms are used for time series forecasting methods to handle non-linear data behavior.

The problem in the AR, ARMA, ARIMA model is the lack of seasonality, which can be addressed in a generalized version of ARIMA model called seasonal ARIMA(SARIMA) and LSTM is resistant to noise (i.e fluctuation of the inputs that are random or irrelevant to predicting a correct output) and the system can be trainable (in reasonable time). From the traditional methods, SARIMA is the most common method used for time series forecasting applications that is proposed in this thesis work. On the other hand, LSTM is proposed and used in this work because it captures long-term dependency and complexity that can also increase the accuracy of the model. Details of these models are discussed in the subsequent sections.

The method of predictive analysis, based on Deep learning (LSTM) and Statistical model (SARIMA), was chosen for the Big Data analysis of traffic data. Training, validation, and testing of machine learning models were performed in the deep learning model and statistical models.

CHAPTER THREE: RESEARCH METHODOLOGY

The dissertation research will collect both quantitative data on the ETRE traffic flow forecasting in case of the Addis-Adama branch. Reliable data will be collected from Addis Adama Branch Database which is automatically collected during cash collection. Addis Adama branch has six toll station [ETRE,2014]. Each toll station traffic flow data was collected centrally in Tulu Dimtu Main Toll Station at Head office.

Traffic flow Forecasting using Time series forecasting is a technique for the prediction of events through a sequence of time. The technique is used across many fields of study, from the geology to behavior to economics. The techniques predict future events by analyzing the trends of the past, on the assumption that future trends will hold similar to historical trends. It is an important area of machine and deep learning that is often neglected because there are so many prediction problems that involve a time component. These problems are neglected because it is this time component that makes time series problems more difficult to handle.

Traffic Flow/Volume Forecasting or Time series forecasting can be performed using various methodologies: Neural Networks, Deep learning, SVM (Support Vector Machines), statistical model/Regression Techniques

3.1. Description of the study area

ETRE Addis-Adama branch is the first modern highway in Ethiopia. It was opened for traffic on September 14,2014. It stretches approximately 78kms (53 miles) long with 6 lanes two roads (two divided carriageways, each with 3 lanes). the highway connects the capital city, Addis Ababa from the north to the city of Adama, in the Oromia regional state, in the south. Currently on a typical workday, it serves approximately 26,000 vehicles on average per day. The ITS and Toll Systems become operational as soon as the new high was opened for traffic in 2014. They were installed and integrated under a designed-build contract, by an international contractor, name Chinese Communications Construction Company (CCCC).

The supervising consultant was also a Chinese company by the name Beijing Expressway Supervision Co. Ltd. The ITS field devices (Mainly CCTV, VMS, and Highway information Management System) are installed close to the toll station. Each toll station has one Entry and Exit toll booth. Each toll booth has one dome camera, a one lane controller system on approaches. In addition, Security CCTV cameras are installed in respective to toll lanes for

monitoring toll collection activities. In effect, high information is there at each toll booth for toll collection and traffic flow registration.

3.2. Research design

A research design is simply the framework of the study of different types of research designs experimental type of research design was employed as the main research design for this study to forecast ETRE traffic flow in the case of Addis-Adama branch. This study uses an experimental research design to explain, understanding, forecasting, and controlling the relationship between values by using different models. ETRE traffic flow data was collected. Then the study determines to what extent traffic flow increase occurs in the future. Accuracy measurement is used to measure the traffic flow increase validity. In this study, the statistical model (SARIMAX) and deep learning model (LSTM) was used. The research followed the following methodology.

- Secondary Data collected by ETRE was obtained.
- This data was converted to a form the models can use.
- To plan for a solution, a model architecture was designed and executed.
- The results were evaluated and compared to the actual values that were recorded by the vehicle detection devices (true value). The difference between the values was known as the error.
- If error was too big, the model architecture was changed to try get more accurate model results.
- After multiple iterations of this process, the model that produced the least error was selected for further analysis.

3.3. Research approach

There are two types of research approach according to Kombo (2006) i.e. quantitative approach this technique uses numerical data or data that are quantified and Qualitative approach that uses non-numerical data or data that have not been quantified. In this thesis the researcher uses a quantitative research approach in analyzing the secondary data thar was collected from September ,2015 to December 31.2021.

3.4. Data Collection

The data are gathered at the Traffic Management Center of ETRE. It is collected for over 2555 days from January 1, 2015, to December 31, 2021, on 24/7 day. The tolling system operates 24 hours per 7 days from 1:00 AM to 29:59:59 PM seven days a week. In this study, a time series of Exit traffic volume data is considered for understanding the exit traffic volume pattern and forecasting future exit traffic volume.

3.5. Data preparation

In the first part of this research, data preparation was performed to obtain files that are suitable for Big Data analysis in the selected software tool. The researcher transformed source files which is found in SQL server to the favorable files suitable for applying SARIMAX and LSTM for using programs written in the Python programming language. Then the source files from SQLsql server were converted to CSV (Microsoft Excel Spreadsheets) files, and the files used directly in the Python were CSV (Comma-Separated Values) files. In python, jupyter notebook "import matplotlib. pyplot" is used for Data visualization.

Moreover, the dataset was prepared into train and test sets (Figure 8). We fit the model on 80% training data and then we evaluate the model on the remaining 20% of test data.

Figure 8: Train and Test model

3.6. Data preprocessing

Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size and their likely origin from multiple, heterogeneous sources. Low-quality data will lead to low-quality mining results [N. Hung (2008), O. Maimon (1999)]. Hence, Data preprocessing is required to have a data set that is suitable for analysis.

Preprocessing of the data in preparation for the prediction can involve data cleaning to reduce noise or handle missing values, relevance analysis to remove irrelevant or redundant attributes,

and data transformation, such as generalizing the data to higher-level concepts or normalizing the data [N. Hung (2008].

The purpose of data preprocessing is to clean selected data for better quality. Data quality is a multifaceted issue that represents one of the biggest challenges for data mining. It refers to the accuracy and completeness of the data. Data quality can also be affected by the structure and consistency of the data being analyzed. The presence of duplicate records, the lack of data standards, the timeliness of updates, and human error can significantly impact the effectiveness of the more complex data mining techniques, which are sensitive to understated differences that may exist in the data. To improve data quality, it is sometimes necessary to clean the data, which can involve the removal of duplicate records, and normalizing the values used to represent information in the database [N. Hung (2008), O. Maimon (1999].

Some selected data may have different formats because the data is very huge, and they stored the data in dump files. Then, in order to use the data, it needs to convert in to a suitable format. Because the purpose of the preprocessing stage is to cleanse the data as much as possible and to put it into a format that is suitable for use in later stages.

3.7. Method of data analysis

In this thesis, the ETRE Traffic flow dataset is obtained from Ethiopian Toll Road Enterprise ICT and Information Monitoring Team. Collecting the Traffic flow data from the Addis Adama Traffic management Database which includes Addis Adama toll road Entry and Exit traffic volume per Toll Station was done. The duration of the flow data traffic is seven 7 years from 01-Jan-2015 to 31-Dec-2021 daily-based measurement saved in Comma Separated Values (CSV) file format. In this research, the researcher is going to use previous ETRE actual traffic flow/volume data for the ETRE traffic flow forecasting.

The researchers use existing traffic flow data and train, and test all traffic flow data using time series forecasting model. The trend of traffic data can be seen using deep learning tools and statistical models. The next traffic flow data can be suggested using those tools and future recommendations can be clearly explained.

Time series analysis involves developing models that best capture or describe an observed time series to understand the underlying causes. This field of study seeks the "*why*" behind a time series dataset.

3.8. Dataset

The Dataset was provided from Ethiopian Toll Road Enterprise in case Addis-Adama Branch which consists of the following characteristics:

- PlazaID: The ID of the toll station.
- TransactionOccurTime: The time at which the vehicle exits toll stations
- VehicleType: The data consists of V1, V2, V3, V4, V5, V6, and V7.
- Total_vehicle: total number of vehicles/cars.

For this thesis, the researcher uses only the date and total vehicle to forecast using univariate time series traffic flow forecasting. So, Transaction occur time by date and Total_flow/ vehicle was selected to predict and forecast ETRE Traffic Flow.

3.9. Training and Testing model

The model which is chosen after reviewing other approaches and literature reviews for performing time series forecast is SARIMA Model which stands for Seasonal Auto-Regressive Integrated Moving Average and Long-short memory. After that the researcher converted data from rows and columns format to a time series data which is efficient and beneficial for traffic flow/volume time series forecasting. By converting to time series data, the researcher use date as the dataset index in spite of serial numbers. Figure 8 below shows the visual division of actual traffic flow split into train and test data. Dataset is divided into two parts:

- Training dataset: It will be used for training our deep learning model and Statistical model (LSTM and SARIMA). Training the model by the functions defined by model.fit () in statsmodels.tsa**.**
- Testing dataset: It will be used for testing and validating our model, Testing the model by providing test dataset by model. predict () in statsmodels.tsa.

3.10. Reliability and Validity

The researcher obtained a model for time series that can now be used to produce forecasts and start by comparing predicted values to real values of the time series, which will help to understand the accuracy of the forecasts.

It is important to evaluate forecast accuracy using genuine forecasts. Consequently, the size of the residuals is not a reliable indication of how large true forecast errors are likely to be.

Previous papers select several indicators to measure predictive accuracy performance. In this study, the two classical performance metrics (i.e., RMSE, MAE) are selected to measure the forecasting accuracy of each model.

This experiment uses the root mean square error (RMSE) and the mean absolute error (MAE) as the evaluation index of the model, which are commonly used to measure the accuracy of the variables. The calculation formula is:

$$\text{RMSE}= \sqrt{1/M \sum_{l=1}^{M}(Av_i - P\hat{v}_i)2}$$

$$\text{MAE} == {}^{1}/_{M} \sum_{i=1}^{M}|Av_i - p\hat{v}_i|$$

where M is a total number of observations, AV represents the true value of traffic flow, which is a sample collected, i.e., the sum of the number of vehicles passing in the six-toll station of Tulu Dimtu, Bishoft south and north, Modjo, Adama west A and B, Adama main at each intersection; $p\hat{v}_i$ represents the predicted value of traffic fl, and m represents the number of observed traffic flow samples. The Root Mean Square Error (RMSE), which can also be called standard error, measures the average size of the error. It is the square root of the mean of the squared deviation of the predicted value from the true value. The Mean Absolute Error (MAE) is the average of the absolute errors of the predicted and true values. Therefore, the smaller the values of RMSE and MAE, the better the model.

3.11. Ethical considerations

All data will be collected with the full informed consent of the participants and the collected data will be treated confidentially. All efforts in sampling, collecting, and analyzing research data will be in alignment with the general rules of researching ethical considerations.

CHAPTER FOUR: SYSTEM ARCHITECTURE

In this chapter, the researcher describes the proposed system model for ETRE traffic flow forecasting. Finally, Results and Discussionof this proposed system can be presented in chapter five.

4.1. Overall Traffic Flow Forecasting Architecture

Progressively, the SARIMAX and LSTM model is retrained using different hyper meters with existing traffic to get the accurate traffic volume while improving it through time. A flow chart of the overall traffic forecasting model that was described is depicted in Fig. 9.

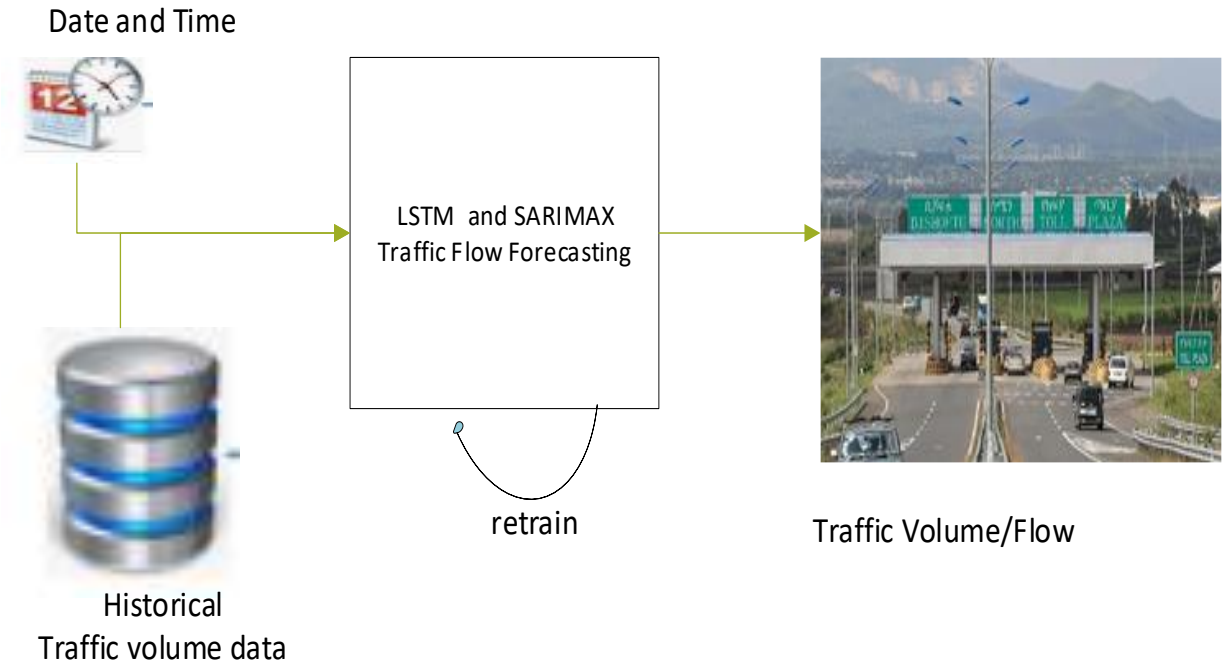


figure 9: Overall traffic flow forecasting architecture

4.2. System Model

The system model presents the steps to be followed for forecasting traffic volume in the ETRE. Figure 10 represents the system model to perform modeling of Traffic flow/volume volumes forecast.

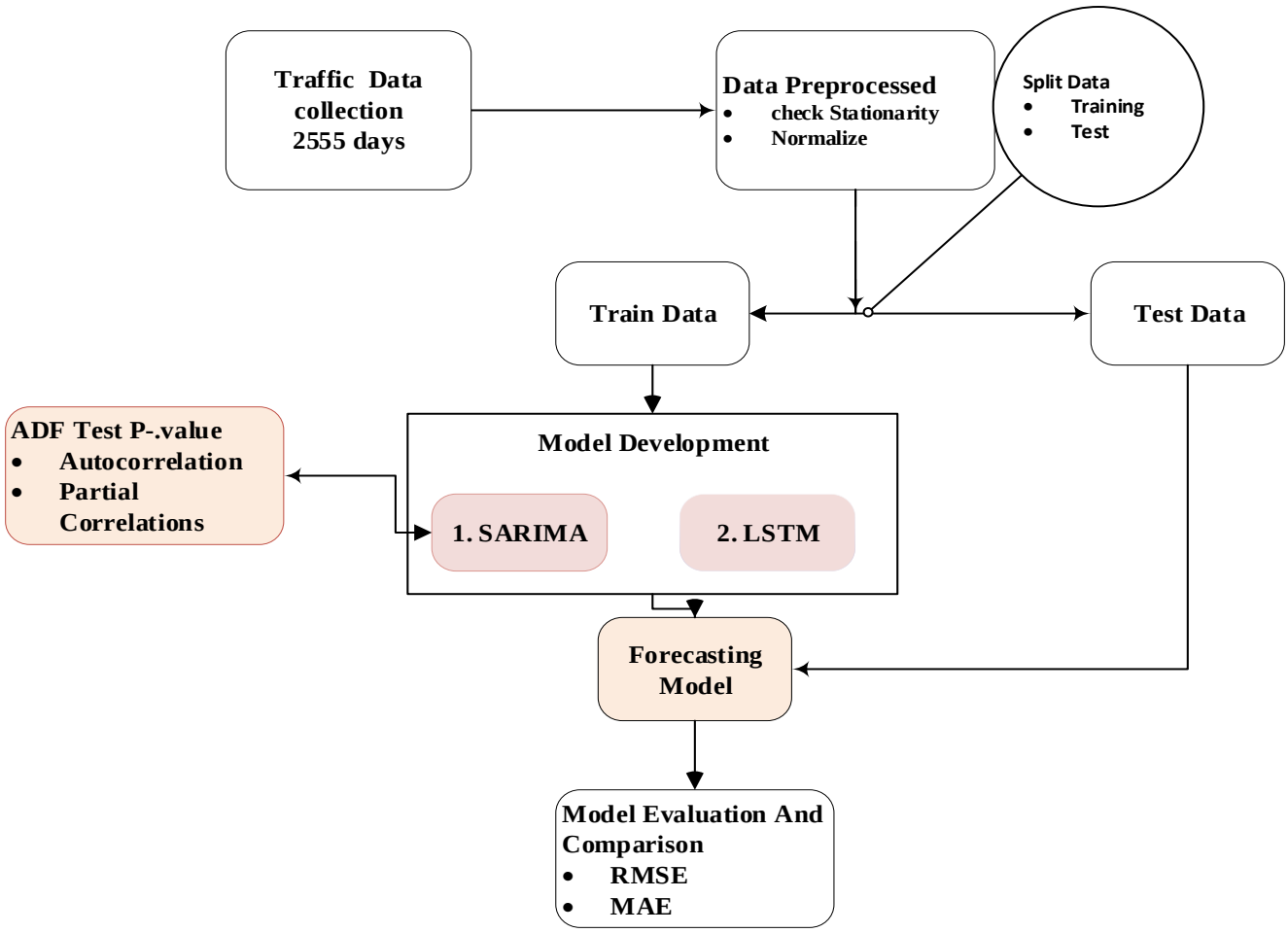


Figure 10: proposed system model

System models are described as follows:

- System model begins with data collection.
- Setting up the system environment
- Then the collected data is pre-processed. Firstly, data is visualized and handling missing values and data normalization are performed. Then, the dataset is divided into two groups: the training set (80%) and the testing set (20%).
- Then next SARIMA and LSTM algorithms are trained using the 80% training dataset.
- The SARIMA and LSTM models are then tested using the 20% testing dataset.
- The models are evaluated based on different performance metrics such as RMSE, and MAE.
- Finally, the one with the minimum prediction error is selected and recommended.

4.2.1. Setting up the system environment

Jupyter Notebook was used as the programming environment. A virtual environment was created to run the experiments for this research. In this virtual environment, the following packages were installed: Tensorflow 1.13.1, Keras 2.2.4-tf, Pandas 0.24.2, Sklearn 0.21.1, Numpy 1.16.3, Matplotlib 3.0, statmodels.

4.2.2. Pre-processing done

Pre-processing of the dataset using Pandas in Jupyter notebook Lab. Pandas is one of the most popular Python libraries for Data Science and Analytics. The researcher used Pandas for handling and performing pre-process of data. The retrieved data is pre-processed before being fed in the SARIMAX and LSTM model. During the Data preprocessing, the collected traffic data has no missing value but it has stationarity issue. Algorithm1and figure 11 below shows how to check data for stationarity

Algorithm 1: Data tranformation and preprocessing

1. Input traffic flow data
2. Read the float32
3. Datetime←convert Datetime
4. ADF←Test
5. If the Data is found Stationary continue to build model
6. ElseIf data is a non-stationary transform it

   Then go to step 4

   End if

7. Go to step 5 repeat

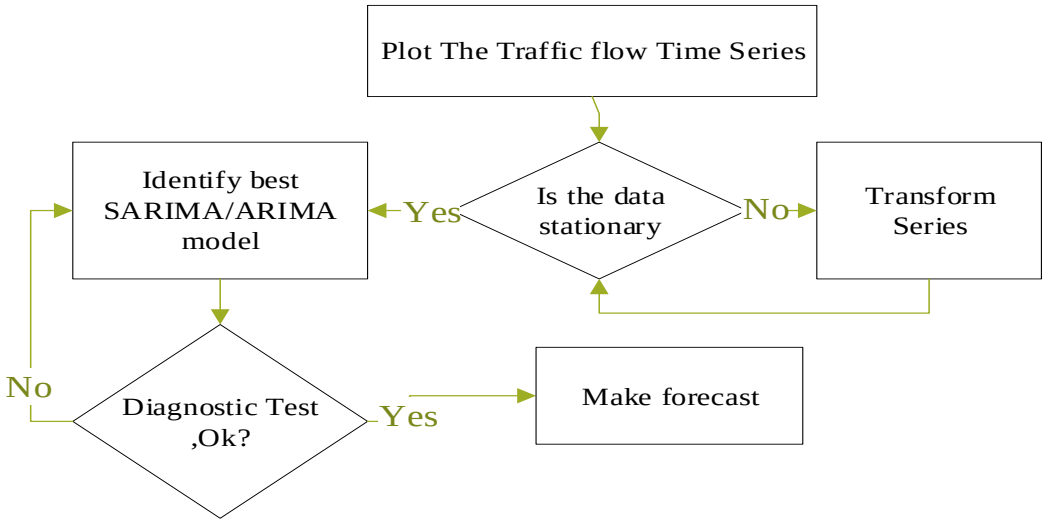


Figure 11:Flow chart of checking Traffic flow Data stationarity

4.2.3. SARIMA Model

There are two ways to evaluate the stationarity of sequences: the first is to judge the stationarity of sequences based on the traits of the sequence diagram and the autocorrelation diagram, and the second is to build test statistics to test hypotheses. The graph test method is an easy and popular way to determine stationarity. The discriminant conclusion has a strong subjective color, which is its drawback. To aid in judgment, it is therefore, employ the statistical test approach. The unit root test is now the statistical test technique for stationarity that is most frequently employed. Figure 12 depicts the SARIMA model's flow chart.

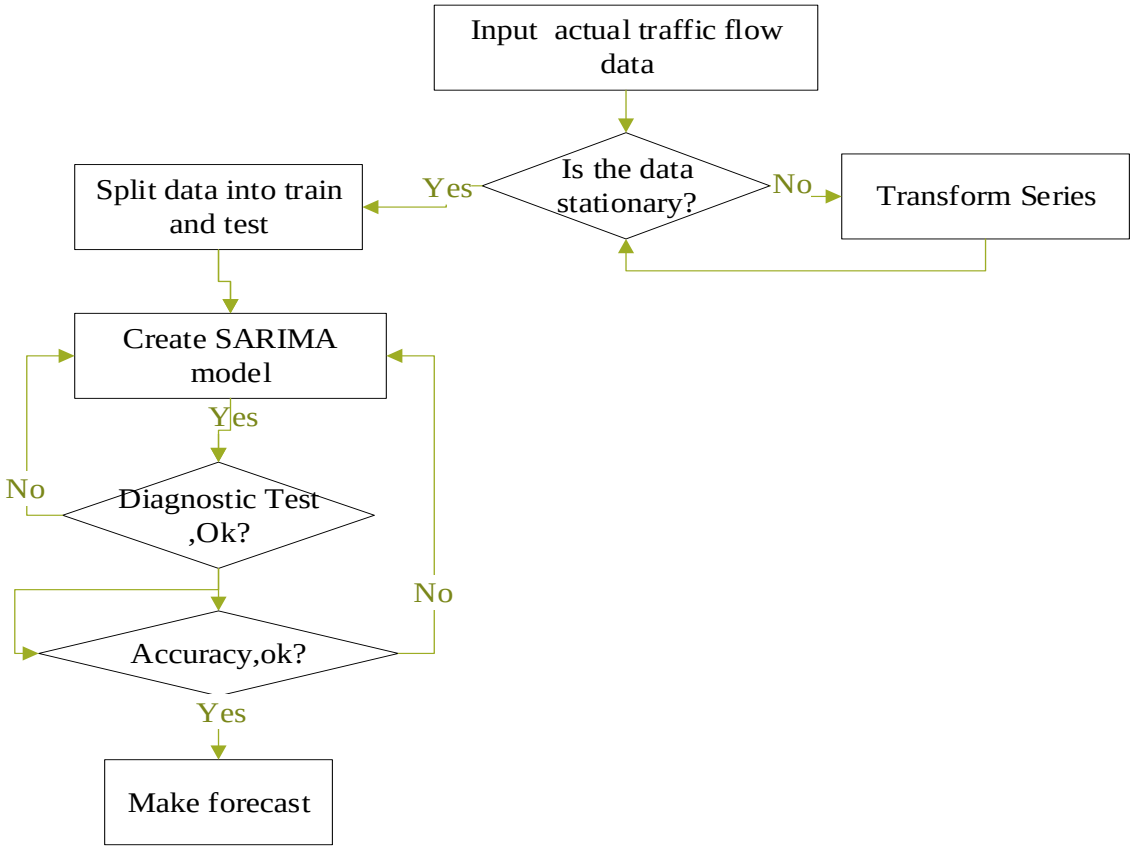


Figure 12: flow chart of Developing SARIMA/ARIMA model

SARIMA is a class of models that captures temporal structures in time series data. SARIMA is a linear regression-based forecasting approach. Therefore, it is best for forecasting a one-step out-of-sample forecast. Here, the algorithm developed performs a multi-step out-of-sample forecast with re-estimation, i.e., each time the model is re-fitted to build the best estimation model [(J. Brownlee, 2017)]. The algorithm, listed in the SARIMA algorithm, takes as input "time series" data set, builds a forecast model, and reports the root-mean-square error of the prediction. The algorithm first splits the given data set into train and test sets, 80% and 20%, respectively (Lines 1-3). It then builds two data structures to hold the accumulatively added training data set at each iteration, "ETRE," and the continuously predicted values for the test data sets, "prediction." As mentioned earlier, a well-known notation typically used in building a SARIMA model is SARIMA (p, d, q) (P, D, Q, S), where: – p is the number of lag observations utilized in training the model (i.e., lag order). – d is the number of times differencing is applied (i.e., degree of differencing). – q is known as the size of the moving average window (i.e., order of moving average). Through Lines 6-12, first the algorithm fits SARIMAX (train, order= (1,0,0), seasonal_order = (0,1,0,81)) model to the test data (Lines 7-8). A value of 0 indicates that the element is not used when fitting the model. More specifically,

an SARIMAX (train, order= (1,0,0), seasonal_order = (0,1,0,81)) indicates that the lag value is set to 1 for autoregression. It uses a different order of 0 to make the time series stationary and finally does not consider any moving average window (i.e., a window with zero sizes. An SARIMAX (train, order= (1,0,0), seasonal_order = (0,1,0,81 forecast model is used as the baseline to model the forecast. This may not be the optimal model, but it is generally a good baseline to build a model, as the researcher explanatory experiments indicated. The algorithm then forecasts the expected value (ETRE) (Line 9), adds theETRE to the prediction data structure (Line 10), and then adds the actual value to the test set for refining and re-fitting the model (Line 12). Finally, having built the prediction and ETRE data structures, the algorithm calculates the RMSE values, the performance metric to assess the accuracy of the prediction and evaluate the forecasts (Lines 14-15)

Algorithm 2: Developed SARIMA Algorithm

#SARIMA

Inputs: Traffic flow Data

Outputs: RMSE of the forecasted data

Output: - forecast Data

Split data into:

80% training and 20% testing data

1. size ← length(series) * 0.80
2. train ← series [0...size]
3. test ← series [size...length(size)]

Data structure preparation

4. ETRE ← train
5. predictions ← empty

Forecast

6. for each t in range(length(test)) do

7. Model=SARIMAX (train, order = (1,0,0), seasonal_order = (0,1,0,81))
8. model fit ← model.fit ()
9. tollroad ← model_fit. forecast ()
10. test_predictions. append(tollroad)
11. observed ← etre_test[t]
12. ETRE. append(observed)
13. end for
14. MSE = mean_squared_error (etre_test, test_predictions)
15. RMSE = sqrt (MSE)
16. Return RMSE

4.2.4. LSTM model

LSTM is a type of recurrent neural network (RNN). RNNs are a powerful type of artificial neural network that can internally maintain the memory of the input. This makes them particularly suited for solving problems involving sequential data like a time series. However, RNNs frequently suffer from a problem called vanishing gradient which leads to the model learning becoming too slow or stopping altogether. LSTMs were created in the 1990s to solve this problem. LSTMs have longer memories and can learn from inputs that are separated from each other by long time lags.

An LSTM has three gates: an input gate that determines whether or not to let the new input in, a forget gate that deletes information that is not important, and an output gate that decides what information to output. These three gates are analog gates based on the sigmoid function which works in the range of 0 to 1. These three sigmoid gates can be seen in Figure 6. A horizontal line that can be seen running through the cell represents the cell state.

Unlike modeling using regressions, in time series datasets there is a sequence of dependence among the input variables. Recurrent Neural Networks are very powerful in handling the dependency among the input variables. LSTM is a type of Recurrent Neural Network (RNN) that can hold and learn from a long sequence of observations.

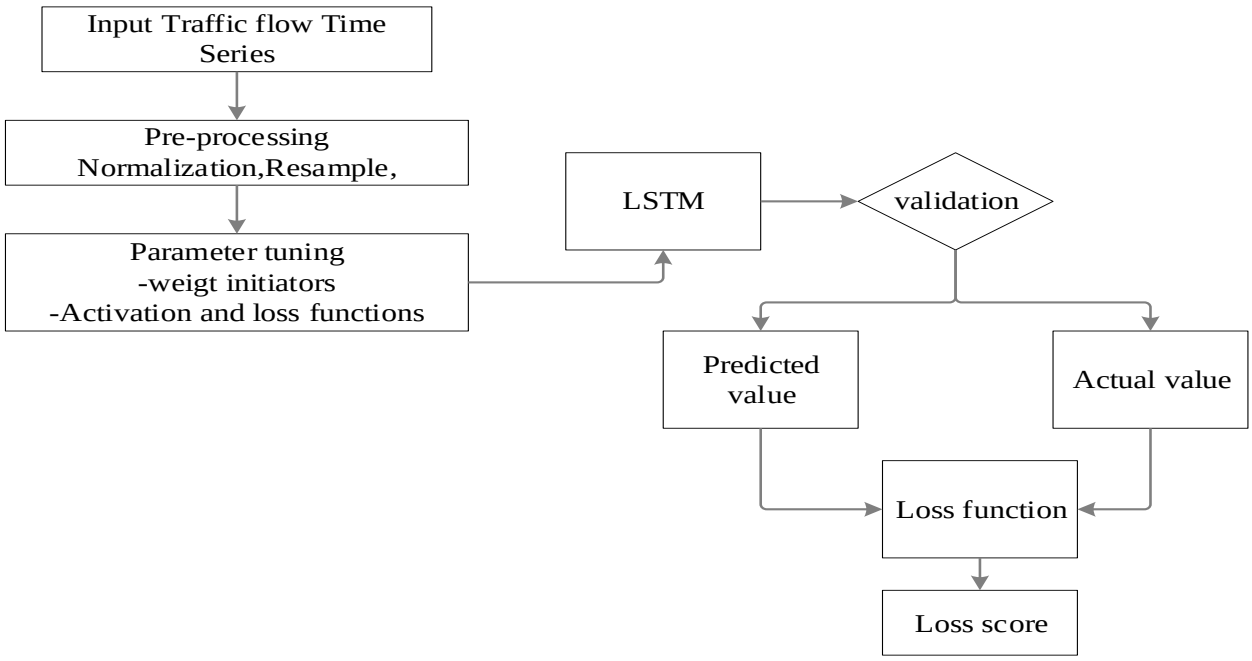


Figure 13: flow chart of LSTM model development

The algorithm developed is a multi-step univariate forecast algorithm [(J. Brownlee, 2016)]. To implement the algorithm, Keras library along with Theano was installed on a cluster of high-performance computing centers. The LSTM algorithm developed is listed in Listing Algorithm 3. To be consistent with the SARIMA algorithm and to have a fair comparison, the algorithm starts with splitting the dataset into 80% training and 20% testing, respectively (Lines 1-3). To ensure the reproduction of the results and replications, it is advised to fix the random number seed. In Line 4, the seed number is fixed to 3 in this research. The algorithm defines a function called "fit lstm" that trains and builds the LSTM model. The function takes the training dataset, the number of epochs, i.e., the number of time a given dataset is fitted to the model, and the number of neurons, i.e., the number of memory units or blocks. Line 8 creates an LSTM hidden later. As soon as the network is built, it must be compiled and parsed to comply with the mathematical notations and conventions used in Theano. When compiling a model, a loss function along with an optimization algorithm must be specified (Line 9). The "mean squared error" and "ADAM" are used as the loss function and the optimization algorithm, respectively. After compilation, it is time to fit the model to the training dataset. Since the network model is stateful, the resetting stage of the network must be controlled specially is more than one epoch (Lines 10 - 13). Also, since the objective is to train an optimized model using earlier stages, it is necessary to set the shuffling parameter to false to improve the learning mechanism. In Line

12, the algorithm resets the internal state of the training and makes it ready for the next iteration, i.e., epoch.

Algorithm 3: The Developed LSTM Algorithm

```
#LSTM

Inputs: Traffic flow data

Outputs: RMSE of the forecasted data

# Split data into:

# 80\% training and 20\% testing data

1. size ← length(series) * 0.80
2. train ← series [0...size]
3. test ← series [size...length(size)]

# Set the random seed to a fixed value

4. set random. seed (3)

# Fit an LSTM model to training data

Procedure fit_lstm (train, epoch, neurons)

5. X ← train
6. y ← train - X
7. model = Sequential ()
8. model.add (LSTM(neurons), stateful=True))
9. model.add (Dense(1))
10. model. compile(loss='mean_squared_error', optimizer='adam')
11. for each i in range(epoch) do
12. model.fit (X, Y, epochs=1, verbose=1, shuffle=False)
13. model. reset_states()
14. end for
```

```
        return model

        # Make a one-step forecast

Procedure forecast_lstm(model, X)

    15. tollroad ← model.predict(X)

return tollroad

    16. epoch ← 500
    17. neurons ← 256
    18. test_predictions ← empty

        # Fit the lstm model

    19. lstm_model = fit_lstm (train, epoch, neurons)

        # Forecast the training dataset

    20. lstm_model. predict(train)

        # Walk-forward validation on the test data

    21. for each i in range(length(test)) do
    22.       #make one-step forecast
    23.     X ← test[i]
    24.   tollroad ← forecast_lstm (lstm_model, X)

        # record forecast

    25.   Test_predictions. append(tollroad)
    26.   expected ← test[i]
    27.   end for
    28. MSE ← mean_squared_error(test_predictions, predictions)
    29. Return (RMSE ← sqrt(MSE))
```

A small function is created in Line 14 to call the LSTM model and predict the next step (one single look-ahead estimation) in the dataset. The number of epochs and the number of neurons is set in Lines 15-16 to 1 and 4, respectively. The operational part of the algorithm starts from Line 18 where an LSTM model is built with given training dataset, a number of epochs and neurons. Furthermore, in Line 19 the forecast is taking place for the training data. Lines 20 - 27 use the built LSTM model to forecast the test dataset, and Lines 28 - 29 report the obtained RMSE values. It is important to note that, for reducing the complexity of the algorithm, some parts of the algorithms are not shown in Box 2 such as dense, batch size, transformation, etc. However, these parts are integral parts of the developed algorithm.

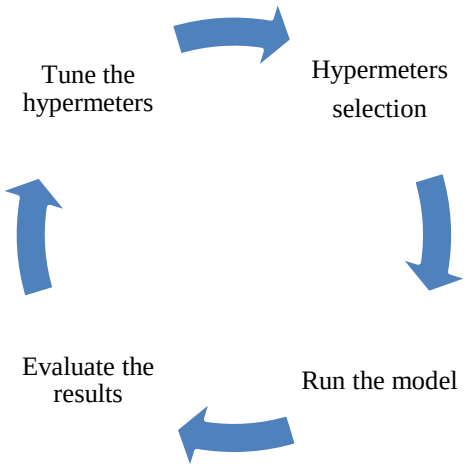


Figure 14: Tuning Hyper-parameters

Figure 14 above shows a typical development cycle for any neural network. For large neural nets, training times can be long so repeatedly running the network and evaluating the results can take days or weeks. For faster development of neural network models, it is important to explore hyper-parameter tuning and develop recommendations that can provide a good initial starting point for the kind of network being developed.

CHAPTER FIVE: RESULTS AND DISCUSSION

Traffic flow prediction and forecasting were developed by SARIMA, and the LSTM model Also, the model parameter was suggested depending on the accuracy of traffic flow/volume forecasted.

5.1. Experimental Results

5.1.1. Brief Overview

Perhaps this is the chapter for which the reader is most eagerly waiting. After gaining reasonable knowledge about time series modeling and forecasting from the previous chapters, the researcher is going to implement them on practical datasets.

In this thesis, till now the researcher considered different time series models, taken from various sources and research works. All the associated programs are written in Python. To judge forecast performances of different methods, the measures MSE, and RMSE are considered. For each dataset, the researcher has presented the obtained results in tabular form. Also, in this chapter, the researcher used the term Forecast Diagram to mean the graph showing the test (actual) and forecasted data points.

5.1.2. Data visualization

Time series plots of raw sample data can provide valuable diagnostics for identifying temporal structures such as trends, cycles, and seasonality that can influence model selection. Figure 15,16 and 17 shows the snapshoot of actual traffic flow collected from ETRE.

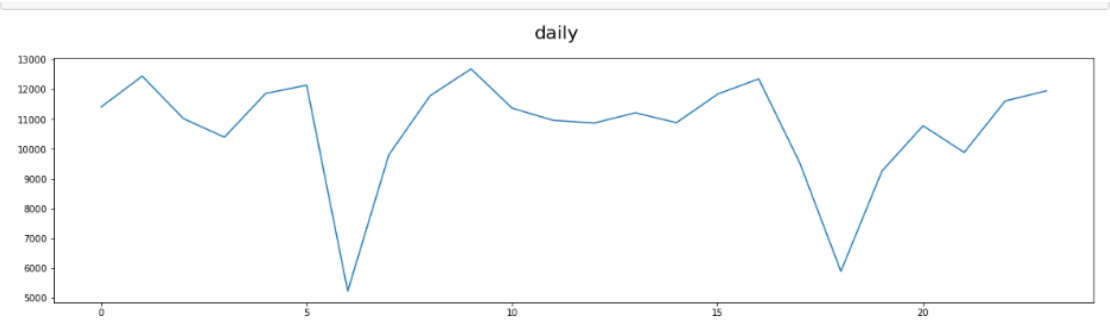


Figure 15: Actual Daily Traffic Flow visualization

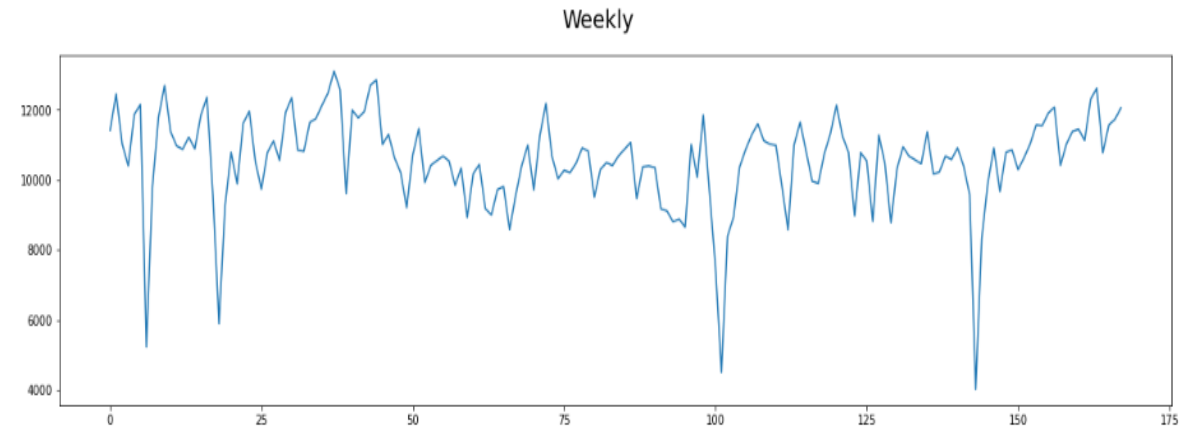


Figure 16: Weekly Actual traffic Flow Visualization

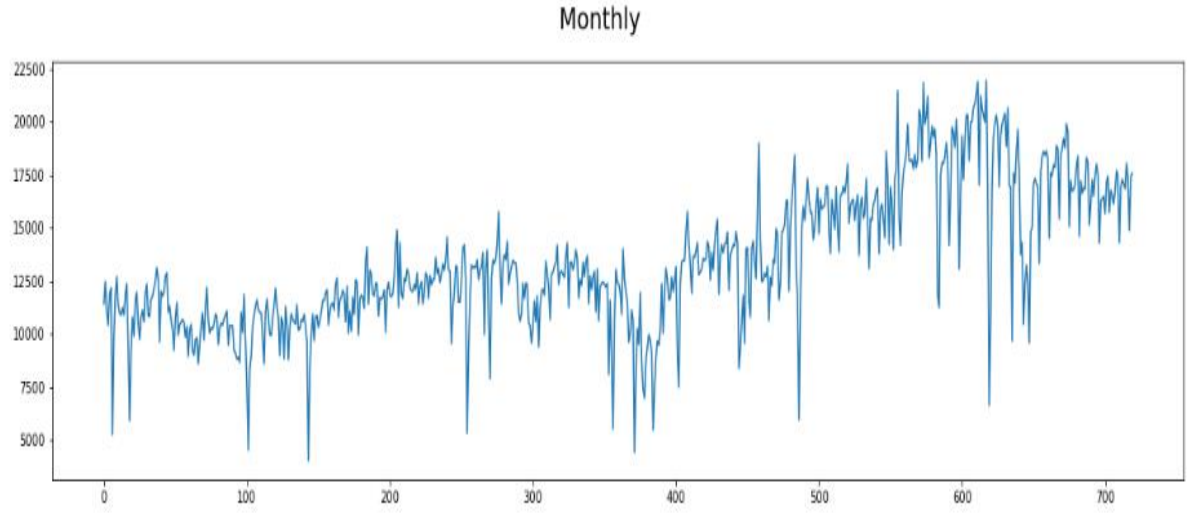


Figure 17: Monthly Actual Traffic flow visualization

5.1.3. Time Series Decomposition

Although decomposition is most frequently used for time series analysis, it may also be utilized as a technique for analysis to help forecasting models better understand the issue. The original time series data for the volume of exit vehicle traffic is decomposed in Figure 18 into its trend, seasonal, and residual components. It analyses characteristics including seasonality and upward and downward patterns. These time units' pattern reveals a seasonal pattern that repeats

every month. The observed pattern of highs and lows, which is routinely repeated, is related to daily seasonal data, which demonstrate seasonality.

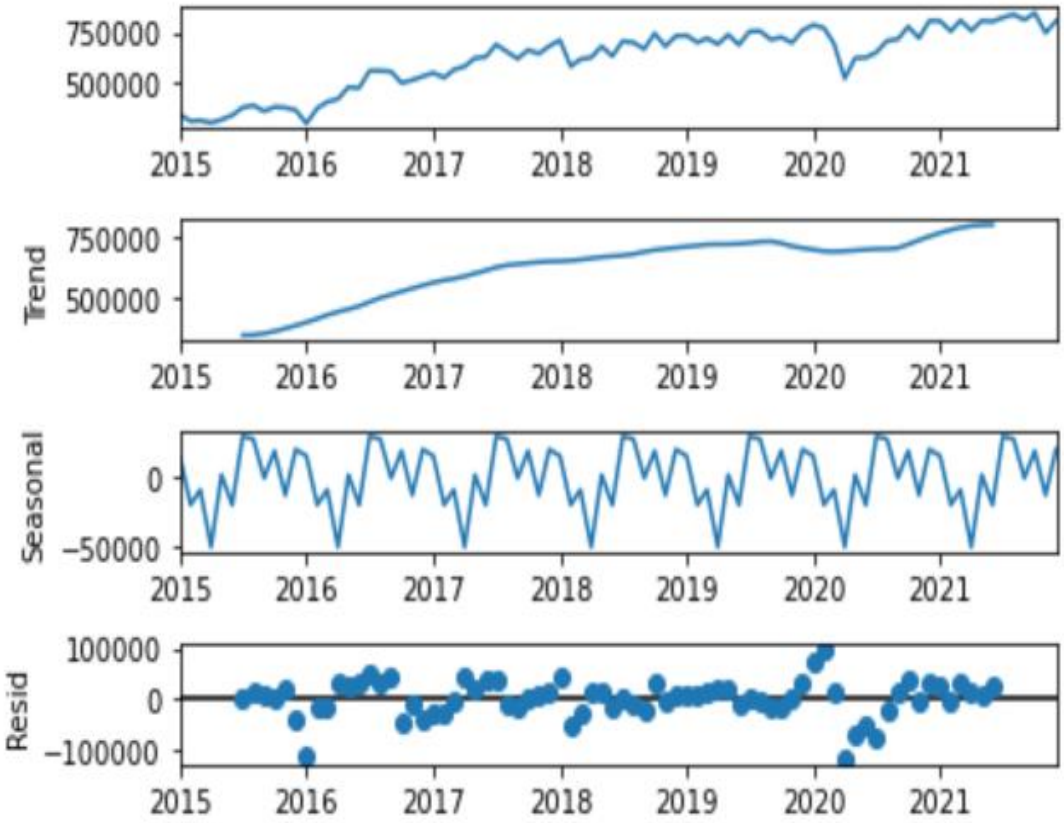


Figure 18: Time series of monthly traffic flow decomposed into trend, seasonality and residuals

5.1.4. ETRE Traffic flow Forecasting Experimental Evaluation and Results

5.1.4.1. Experiment 1: SARIMA Model

Statistical time-series models, such as the Seasonal autoregressive integrated moving average (SARIMA) model, attempt to develop a mathematical model explaining the past behavior of a series and then apply it to forecast future behavior. SARIMA models are an extension of ARIMA that explicitly supports univariate time series data with a seasonal component. The most crucial step in estimating the SARIMA model is determining the values of model seven parameters. To determine the value of SARIMA model parameters using ACF and PACF plots and grid search. The ACF and PACF plots were used as a starting point to help determine a few likely parameters, and then the best parameters were identified using a grid search.

SARIMA is an extension of ARIMA. SARIMA models have been applied to the UTCS [Okutani and Stephanedes,1984] and freeway volume forecasting [Kim and Hobeika ,1993].

SARIMA models rely on a seasonal and non-seasonal series of data. For this study, a simplified SARIMA model was developed based on monthly/daily volumes for each time interval, computed from the Addis-Adama traffic volume in CSV format. The model was developed with seven consecutive years (January 1, 2015 to December 31,2021) of data, with no missing values. Using a statistical software package, and SARIMA (P, D, Q) (P, D, Q, S) process was identified and then the parameters were estimated.

Pandas is a powerful Python data analysis toolkit. It is an open-source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

5.1.4.2. SARIMA Model Parameter Analysis

1. ACF and PCF plot

The ETRE traffic volume dataset for monthly data has a total of 84 observations and for Daily 2555 observations. From monthly data, the first 64 observation are considered for training and the remaining 18'n observation was used for testing. An SARIMAX model of order SARIMAX (1, 1, 0) x (1, 1, 0, 25) is the most parsimonious SARIMA model for this series. Figure 19 and 20 Below have shown the sample ACF and PACF plots for the ETRE traffic volume data series.

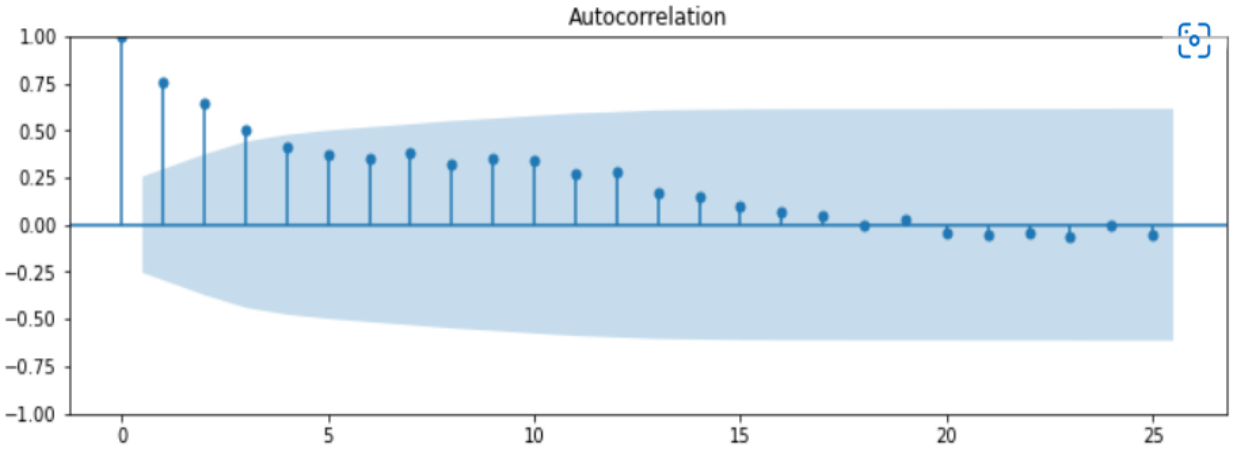


Figure 19 : Autocorrelation of Monthly traffic flow

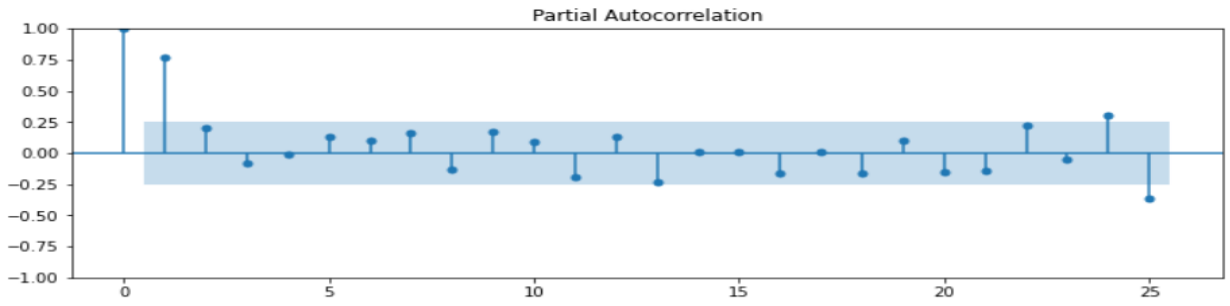


Figure 20: PCF of monthly traffic flow

2. Stationary Test

Test the stationary of the time series through a unit root test, ADF test, and examine its Exit traffic flow for trend and seasonality. A statistically significant test results in the null hypothesis (p-value > 0.05) for this test is that the data is not stationary and the null hypothesis (p-value < 0.05) for this test is that the data is stationary. The statistical test result from the ADF test the p-value was 0.0032. Based on the result the p-value <0.05, so the data is stationary.

3. SARIMA Model selection

The SARIMA algorithm is applied to forecast the Exit traffic flow data. During the model selection, the researcher uses red color for prediction and the grid search method is used for hyperparameter optimization and determining the appropriate forecasting model parameters. The initial value of SARIMA model parameters is determined through ACF and PACF plots, evaluation of in sample forecast of SARIMA (1, 1, 0) (1, 1, 1)80. With a starting point of (1, 1, 1) (1, 1, 1)80 a grid search was set up to test several different parameter combinations. The models were evaluated using the AIC criterion. 1 display some of the model grid search results that show the models with the lowest AIC score values for monthly traffic flow prediction. So, the final SARIMA model parameters are used in SARIMA (1, 1, 1) (1, 0, 1)81.

Non-seasonal parameter	Seasonal parameter	Model selection criterion	Total fit time
p, d, q	P, D, Q, s	AIC value	Sec
0, 0, 0	0, 1, 0, 80	99.613	0.17 sec
0, 0, 0	0, 1, 0, 81	74.7	0.09 sec
12, 1, 1	0, 1, 1, 12	456.966	

Table 1 : SARIMA Model Parameter and selection for monthly traffic flow prediction

4. Diagnosis check

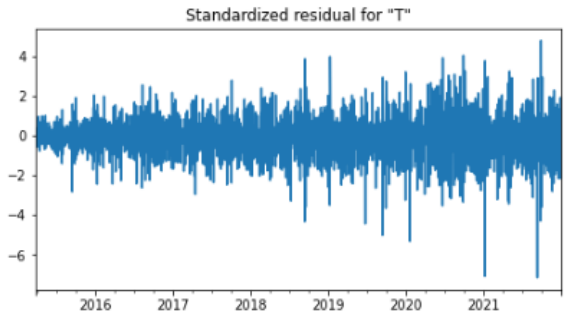
A diagnosis check is used to validate the selected model's forecast accuracy. The AIC value of SARIMA (1, 0, 0) (1, 0, 1)81 is the lowest as shown in figure 21 depicts the diagnostic check of the chosen SARIMA model summary, which shows that the p-value is less than the significant value (0.05), so this indicates that the forecast accuracy of the chosen model is well. The table also shows the SARIMA parameter value of lags.

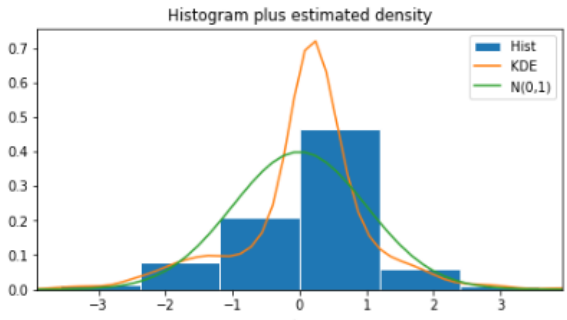
Figure 21: The diagnostics test results of the SARIMA (1, 0, 0) (1, 0, 1,)81 model.

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.9926	0.002	417.074	0.000	0.988	0.997
ar.S.L81	-0.5142	0.123	-4.198	0.000	-0.754	-0.274
ma.S.L81	0.5991	0.116	5.163	0.000	0.372	0.827
sigma2	6.143e+06	1.04e-08	5.93e+14	0.000	6.14e+06	6.14e+06

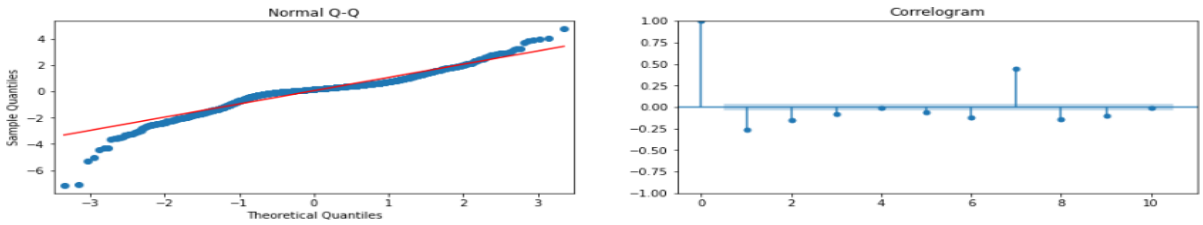
5. Residual Plot

Figure 22, (a), depicts the residuals over time. The findings imply that the residuals have no discernible seasonality and appear to be white noise. Similarly, the autocorrelation in Figure 22 (d) indicates that the residuals of the original data have a low correlation with the lagged data. According to Figure 22(b), the Kernel Density Estimation (KDE) (red curve) is nearly overlapping with the N (0, 1) (green curve). The results indicate that the residual has a normal distribution, with a mean of 0 and a standard deviation of 1. The red line in Figure 22 (c) represents normally distributed traffic volumes at exit toll station with a mean of zero and a standard deviation of one, while the blue dots represent residuals. In general, the Q-Q plot shows residuals that are normally distributed, indicating that the chosen model fits well and can be used to forecast future values.





(a) (b)



(c) (d)

Figure 22 : Plot of residual: (a) Residuals over time; (b) Distribution histogram; (c) Q-Q plot and (d) Autocorrelation

5.1.4.3. Experiment 2: LSTM Model Evaluation

LSTM is an RNN that can save and learn from long-distance tracking. The advanced LSTM model is a consistent multi-step prediction algorithm. The researcher first set the training_size to 2044 data points past values used to train the exit traffic volume. The traffic flow algorithm is used in Python, Keras, and Tensorflow as a backend.

```
Model: "sequential_6"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 lstm_6 (LSTM)                (None, 256)               1696768

 dense_6 (Dense)              (None, 1)                 257

=================================================================
Total params: 1,697,025
Trainable params: 1,697,025
Non-trainable params: 0
```

Figure 23 :LSTM Architecture

Hyperparameter values are used to control the learning process. The LSTM model was tested with various hyperparameters in this case. To train and predict the target with great accuracy, it is important to choose the right hyperparameter combination. LSTM properties are selected and tested, as is the model with the best hyperparameters. The hidden layers of LSTM are limited to 3 and the number of epochs and neurons is set to 500 and 1000 respectively. These are some of the hyperparameters to choose from and can affect the tradeoff between predictive accuracy and training time. A large number of layers can improve predictive accuracy. Based on the relationship between the number of previously recognized values and the accuracy of the prediction, which determines the amount of information the network needs to remember and use. We use the first 2044 days of data for training and 511 days for testing/verification.

The “Sequence” and “ADAM” are used as a function of the function and the algorithm for successive sequences. Table 2 shows the list of selected hyperparameters and their values in this research.

Hyperparameters	Values	Remark
Hidden Layer	3	
Number of Neurons	256	
Number of Epoch	1000	
Learning rate	0.001	
Bach size	54	
Verbose	1	
Optimizer	Adam	
Loss	MeanSquaredError	
Metrics	RootMeanSquaredError	

Table 2 Hyperparameters and values for LSTM model

5.1.5. LSTM and SARIMAX Models Evaluation Results

Evaluation is one key point in any Deep learning process. It serves two purposes: the prediction of how well the final model will work in the future and an integral part of many learning methods, which help to find the model that best represents the training data. In the series of experiments, evaluation of models is done based on performance/accuracy of models and confusion matrix, discussion with the domain expert and based on the soundness of the rules generated.

The forecasting performance of a model can be examined by the standardized statistical tools such as root mean square error (RMSE), mean absolute error (MAE). Table 3 and 4 shows LSTM and SARIMAX accuary output.

Table 3: In-sample forecast accuracy measures results for SARIMAX ((12,1,4) (0,1,1,7)) model

Acuracy measures	values
Root Mean Square Error (RMSE)	1800
Mean Average Error (MAE)	1151
LMSE	0.21

Table 4: In-sample forecast accuracy measures Results for LSTM model

Acuracy measures					
Train values			Test values		
RMSE	MAE	LRMSE	RMSE	MAE	LRMSE
974	752	0.04	0.0007	0.0002	2.82
1278	846	0.06	2098	1682	0.08

Figure 24 and 25 are snapshots of the transition between training and testing data. The 'Predictions' line represent the model output after training with the training dataset and the testing dataset being input into this model. The graphs illustrate the model predictions compared to the testing dataset which was reserved to check the accuracy of the model.

Figure 24 : Actual and forecast Traffic flow using LSTM

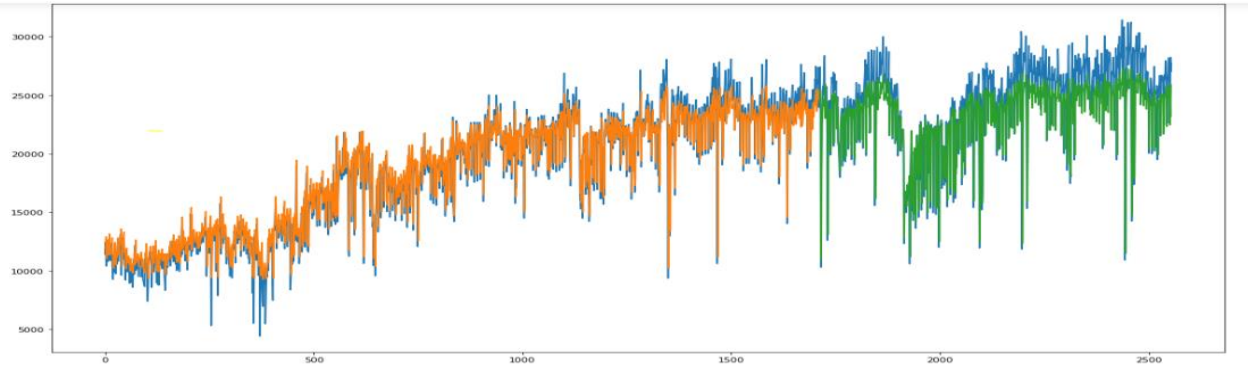


Figure 25: Actual and predicted Traffic flow prediction using SARIMA

Figure 26 and 27 are snapshots of the out of sample forecast of future daily traffic flow of ETRE using LSTM amd. The forecasted data was shown on appendix 2 and its graphical visualization snapshot was shown in appendix 3 and 4

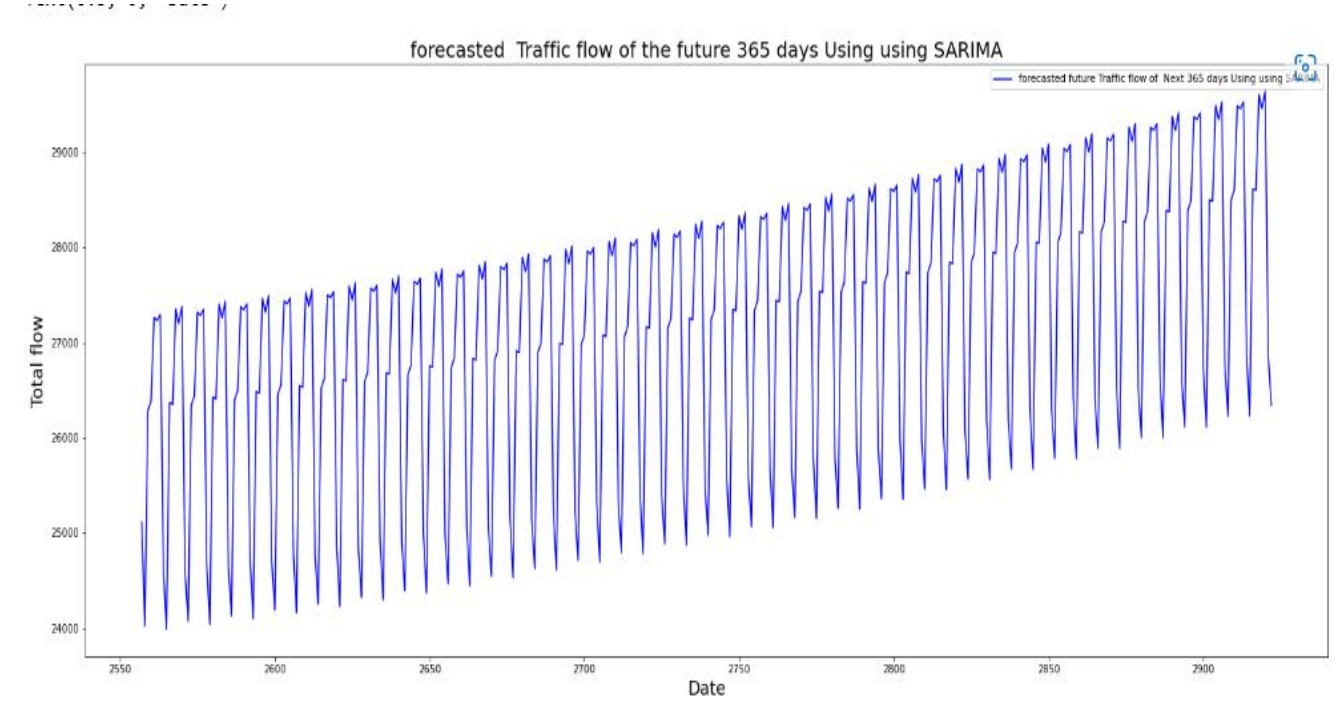


Figure 26: SARIMA out of sample forecast of ETRE traffic flow forecasting

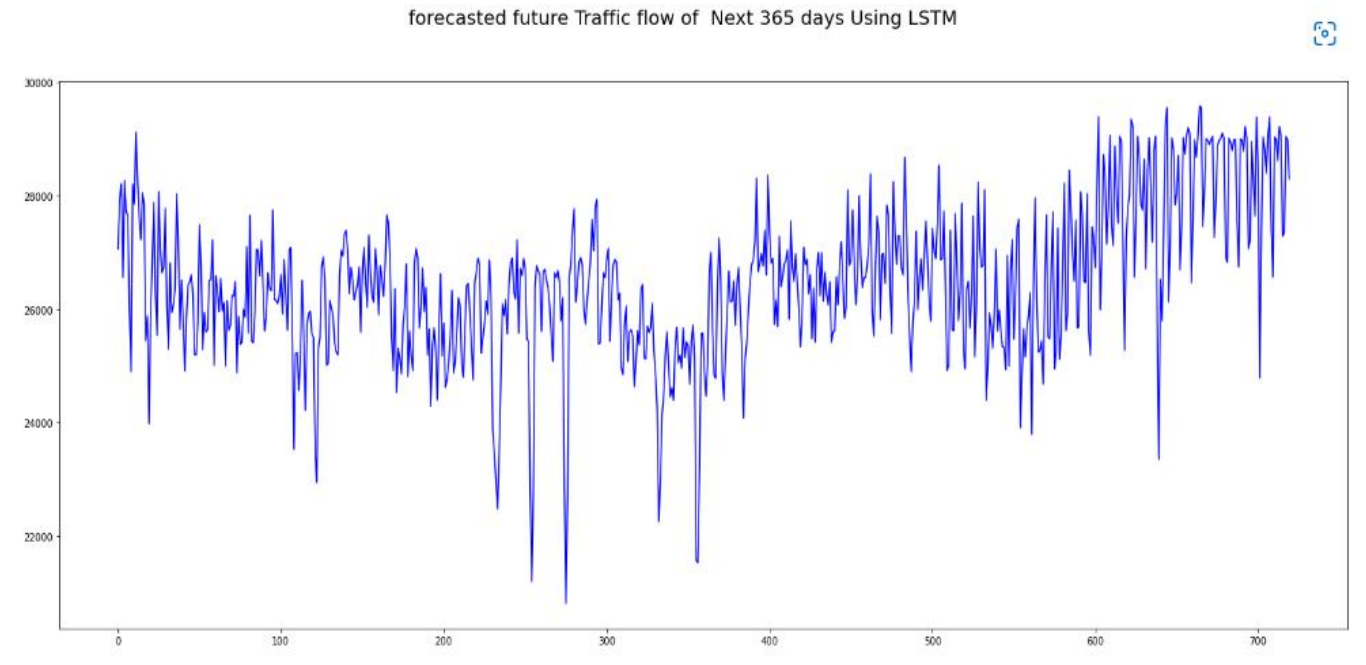


Figure 27: Out of sample forecast future traffic flow forecasted using LSTM

CHAPTER SIX: CONCLUSION AND RECOMMENDATION

6.1. Conclusion

Today, Traffic flow are becoming more and more critical to many medium and large transportation especially on toll roads. In the toll road, the effective management of traffic volume is a major problem due to its flexible behavior traffic congestion. That means it is not easy to predict Traffic volume that the toll road to provide appropriate services including staff planning and technology migration from manual to automatic number plate recognition and contactless technology. Therefore, an algorithm that better understands complex and unrelated relationships of traffic volume is needed to improve prediction accuracy.

The purpose of this study was to develop a model that predicts the volume of traffic exit using historical data from Ethiopian Toll Road Enterprise. The researcher spent for months data record to build a model. It is divided into 80% training set and 20% test set. We used two consistent time series strategies, namely SARIMA and LSTM to create a model and tested to predict the volume volume of traffic volume at exit toll station.

Test results show that the LSTM model has error metrics in RMSE, lower MSE compared to the SARIMA model. The LSTM model includes many features in traffic flow forecasting that are not included in the SARIMA model. Therefore, the overall result suggests that the LSTM method may be considered an effective way to predict traffic flow/volume in ETRE to show interim patterns and improve resource allocation to reduce the traffic flow congestion and plan for the future traffic flow.

6.2. Recommendation

We recommend the Ethiopian Toll Road Enterpsie to use this model to predict traffic flow to reduce traffic congestion by providing staffing and technical improvements to streamline staffing and increase customer satisfaction. It may also help future researchers to:

- Develop a software for Traffic flow live data forecasting using LSTM model
- Further research into the multivariate time series to opt out of traffic flow to improve the prediction of certain types of toll roads.

References

A. A. Adebiyi, A. O. (vol. 2014,). *"Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction,"* , pp. 9–11,.

A. Simroth and H. Zähle. (Mar. 2011.). "Travel time prediction using floating car data applied to logistics planning," . *IEEE Trans. Intell. Transp. Syst., vol. 12,no. 1, pp. 243–253,* .

Agrawal, R. A. (n.d.). In *"An Introductory Study on Time Series Modeling and Forecasting ."*.

B. De Constance and B. De Constance. (n.d.). "Jenkins methodology." .

B. Sennaroglu and G. Polat. (no. JUL, vol. 2017,). "Time series forecasting for a call centre," g . *Proc. Int. Conf. Ind. Eng. Oper. Mana g*, pp. 464–468, 2017.

Brian L. Smith a, *. B. ((2002)). Transportation Research Part C 10. *Comparison of parametric and nonparametric models for traffic flow forecasting*, 303–321.

Centiner, B. G. (2010). A neural network based traffic-flow prediction model. *Mathematical and Computational Applications*, , 15, 269-278. .

D. K. Barrow. (2016). *"Forecasting intraday call arrivals using the seasonal moving average method," , doi: 10.1016/j.jbusres.2016.06.016.* , J.Bus. Res., vol. 69, no. 12, pp. 6088–6096, .

Davis, P. J. (n.d.). In *Introduction to Time Series and Forecasting , Second Edition Springer Texts in Statistics.* .

E. R. Pasupathy, S.-H. K. (2013.). USING SIMULATION TO EVALUATE CALL FORECASTING ALGORITHMS FOR INBOUND CALL CENTER Guilherme,. pp. 1132–1139, .

Ethiopian Toll Road Enterprise. (2012). *ETRE 5 years Strategic Planning 2012-2017.* Addis Ababa: Planning and Report Team.

ETRE. (2022, january 31). Ethiopian Toll Road Enterprise monthly Report. p. 35.

ETRE. (2022). *Ethiopian Toll Road Enterprise,.* Management.

(2014, july 24). *Federal negarith Gazet of federal democratic republic of ethiopia.* Addis Ababa: NEGARIT GAZETA.

freewaymanagement. (2014). Retrieved from https://ops.fhwa.dot.gov/freewaymgmt/publications/frwy_mgmt_handbook/chapter15_02.htm.

H. Palangi, R. W. (pp. 4504–4518, 2016,). "Distributed Compressive Sensing: A Deep Learning Approach," , doi:10.1109/TSP.2016.2557301. . *IEEE Trans. Signal Process.,*, vol. 64, no. 17, .

H. Shen and J. Z. Huang. (January 2003, 2014.). Forecasting Arrivals to a Telephone Call Center,.

H. Zou and Y. Yang. (2004). *"Combining time series models for forecasting," ,doi:10.1016/S0169-2070(03)00004-9.*, vol. 20, pp. 69–84,.

H. Zou and Y. Yang. (2004). *“Combining time series models for forecasting,”,doi:10.1016/S0169-2070(03)00004-9.*, vol. 20, pp. 69–84,.

H. Zou and Y. Yang. (2004). “Combining time series models for forecasting,” ,doi: 10.1016/S0169-2070(03)00004-9. . vol. 20, pp. 69–84, .

I. Khandelwal, R. A. (2015). *“Time Series Forecasting using Hybrid ARIMA and ANN Models based on DWT Decomposition,” , doi: 10.1016/j.procs.2015.04.167.* , vol. 48, no. Iccc, pp. 173–179, .

I. Khandelwal, R. A. (2015,). “Time series forecasting using hybrid arima and ann models based on DWT Decomposition,”, doi: 10.1016/j.procs.2015.04.167. . *in Procedia Computer Science,*, vol. 48, no. C, pp.173–179.

I. Khandelwal, R. A. (2015,). “Time series forecasting using hybrid arima and ann models based on DWT Decomposition,. *” in Procedia Computer Science, vol. 48, no. C,doi: 10.1016/j.procs.2015.04.167.*, pp.173–179,.

ITS. (2022). Retrieved from https://rno-its.piarc.org/en/network-operations-its-road-safety-speed-management/use-vms.

J. Bayer and C. Osendorfer. (2014). “Learning Stochastic Recurrent Networks,” ,http://arxiv.org/abs/1411.7610. pp. 1–9, ,.

J. Bayer and C. Osendorfer. (2014,). “Learning Stochastic Recurrent Networks,” ,[Online]. Available: http://arxiv.org/abs/1411.7610. . pp. 1–9, .

J. Bayer and C. Osendorfer. (pp. 1–9, 2014). Retrieved from http://arxiv.org/abs/1411.7610.

J. Brownlee. (2016). “Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras,” https://machinelearningmastery.com/ime-series-prediction-lstm-recurrent-neural-networks-python-keras/,.

J. Brownlee. (2017). *” https://machinelearningmastery.com/*.

J. W. Taylor and J. W. Taylor. (no. September 2015, 2008). A Comparison of Univariate Time Series Methods for Forecasting Intraday Arrivals at a Call Center,” , doi: 10.1287/mnsc.1070.0786. *Forecasting Intraday Arrivals at a Call Center*.

KapschTrafficCom. (n.d.). Retrieved from https://www.kapsch.net/en/solutions/tolling.

Lippi, M., Bertini, M., & Frasconi. (P.2013). short-term traffic flow forecasting:An experimental comparison of time-series analysis and supervised learning. *IEEE Transactions on Intelligent Transportation System*, 14(2):871-882.

M. Milenković, L. Š. (no. April, 2016,). *“SARIMA modelling approach for railway passenger flow forecasting, doi:10.3846/16484142.2016.1139623.* , *”* vol. 4142, .

N. Hung. (2008.). "Data Cleaning and Data Preprocessing Techniques,". *Academic Press,* .

O. Maimon. (1999). "Introduction to Data Mining and Knowledge Discovery",. *Two Crows Corporation,* .

P. J. Brockwell and R. A. Davis. (n.d.). Introduction to Time Series and Forecasting , Second Edition Springer Texts in Statistics. . .

P. P. Dabral and M. Z. Murry. (2017,). "Modelling and Forecasting of Rainfall Time Series Using SARIMA," doi: 10.1007/s40710-017-0226-y.

P. Ross. (1982,). "Exponential filtering of traffic data," . *in Proc. Transp. Res. Board, Washington, DC, vol. 869,*, pp. 43–49.

Rohit Galani1, *. D. (2020). Prediction of Traffic using Economic Attributes Volume 5, Issue 1.

S. Innamaa. (Apr. 2006.). "Effect of monitoring system structure on short-term prediction of highway travel time," . *Transp. Planning Technol., vol. 29, no. 2,*, pp. 125–140.

S. Siami-namini and N. Tavakoli. (2018). *"A Comparison of ARIMA and LSTM in Forecasting Time Series," , doi: 10.1109/ICMLA.2018.00227.* , pp. 1394–1401,.

Sun, S. L. (2006). A Bayesian network approach to traffic flow forecasting. *IEEE Transactions on intelligent transportation systems*, 7.1: 124-132.

Thabassum, S. (2013;). Impact of State-Wise Vehicle Contribution on Traffic Growth Rates for National Highways. *International Journal for Engineering Research and Technology (IJERT).*, 2(10): 2101-2106.

Theja, V. (2010). Short Term Prediction of Traffic Parameters Using Support Vector Machines Technique. *Proceedings of third international conference on Emerging trends in Engineering and Technology*, 70-75.

Traffic Flow Theory. (2022). Retrieved from https://www.webpages.uidaho.edu/niatt_labmanual/chapters/trafficflowtheory/theoryandconcepts/TrafficFlowParameters.htm.

U. Yolcu, E. E. (pp. 1340–1347, 2013). "A new linear & nonlinear arti fi cial neural network model for time series forecasting," , , doi:10.1016/j.dss.2012.12.006. . *Decis. Support Syst.*, vol. 54, no. 3, .

V. B. Arem, H. R. (Mar. 1997). "Recent advances and applications in the field of short-term traffic forecasting. *" Int. J. Forecasting, vol. 13, no. 1,* , pp. 1–12,.

W. Bao, J. Y. (2017). "A deep learning framework for financial time series using stacked autoencoders and long-short term memory,". *doi:10.1371/journal.pone.0180944.*, vol. 12, no. 7, pp. 1–24,.

W. Bao, J. Y. (2017,). *A deep learning framework for financial time series using stacked autoencoders and long-short term memory," PLoS One, doi:10.1371/journal.pone.0180944.*, vol. 12,no. 7, pp. 1–24,.

W. Bao, J. Y. (pp. 1–24, 2017,). *A deep learning framework for financial time series using stacked autoencoders and long-short term memory," PLoS One,doi:10.1371/journal.pone.0180944.*, , vol. 12, no. 7,.

WEI, M. (01,2013). *ROAD ARRANGEMENT.* Addis Ababa: China communication Construction Company LTD.

wikipedia. (n.d.). Retrieved from https://en.wikipedia.org/wiki/Toll_road.

Zhao, Z., Chen, W., Wu, X., Chen, P., & Liu, J. (2017). LSTM network: A deep learning approach for short-term traffic forecast. *IET Intell. Transp. Syst*, 11, 68–75.

Appendix

1. General proposed system Architecture

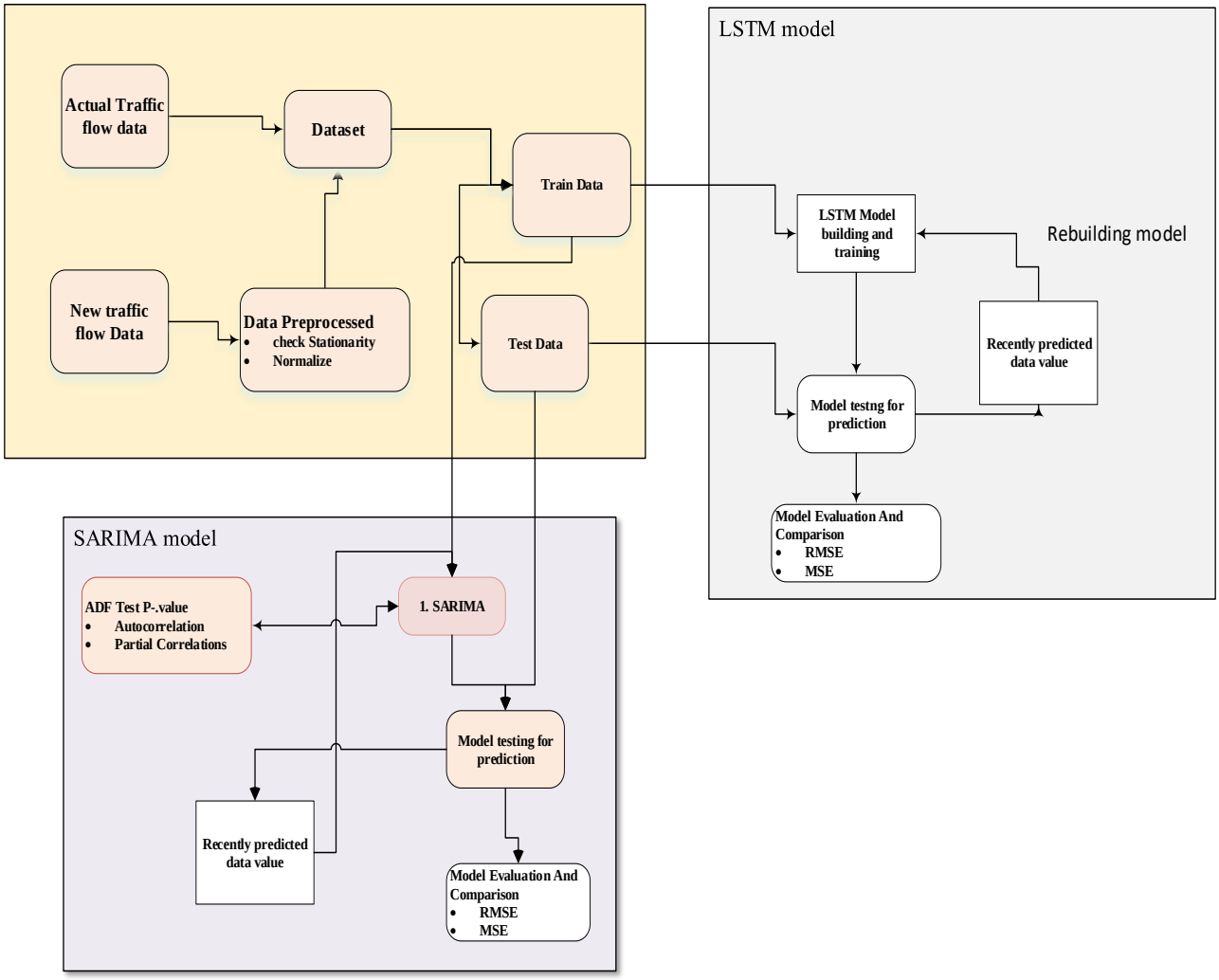


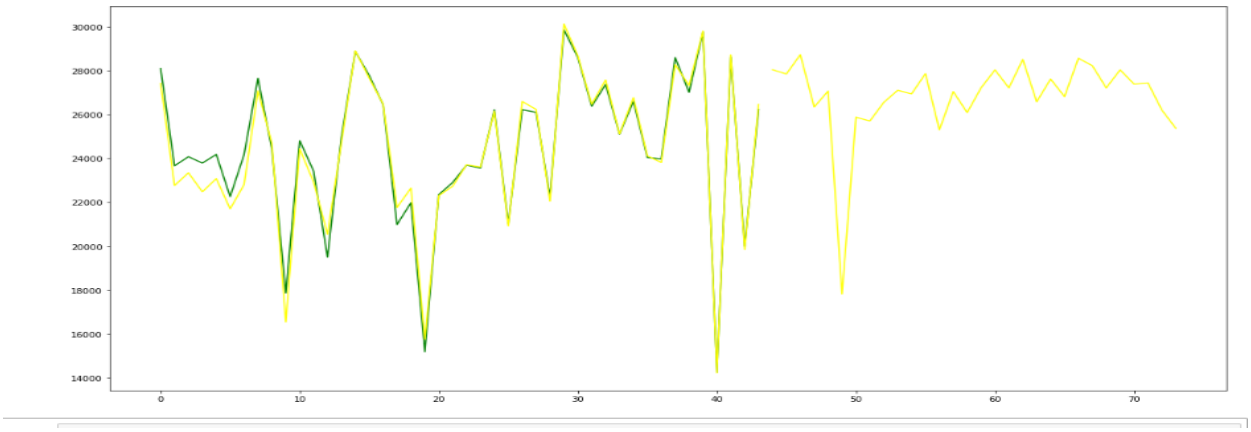
Figure 28:General system architecture used in this research

2. Out of sample Forecast Data

Timestamp (Date)	Forecast Traffic flow Using SARIMAX (0, 0, 1) x (0, 1, 0], 81)	Forected Traffic flow using LSTM
1.	26964	26975
2.	26766	27620
3.	26888	27662
4.	26454	23779
5.	21871	27013
6.	25913	27183
7.	26910	27038

8.	26497	26184
9.	26967	24692
10.	28111	21581
11.	27829	22995
12.	22615	27324
13.	25790	27464
14.	26247	27592
15.	26814	27644
16.	27129	26033
17.	28279	27963
18.	28358	24076
19.	22770	22878
20.	26234	21893
21.	23740	20897
22.	28407	21496
23.	27178	21561
24.	27865	23264
25.	28223	21729
26.	27121	24902
27.	26924	26845
28.	27045	26534
29.	26611	26945
30.	22028	27098
31.	26070	25970

3. Snapshoot of in-sample and out-sample forecast in graph SARIMA



Graph 1 Graphical visualization of Daily actual, train, Forecast Data using LSTM

4. Snapshoot of in-sample and out-sample forecast in graph SARIMA

```
etre['totalflow'].plot(figsize=(12,8),legend=True)
forecast.plot(legend=True,color='red')
```

```
<AxesSubplot:xlabel='Month'>
```

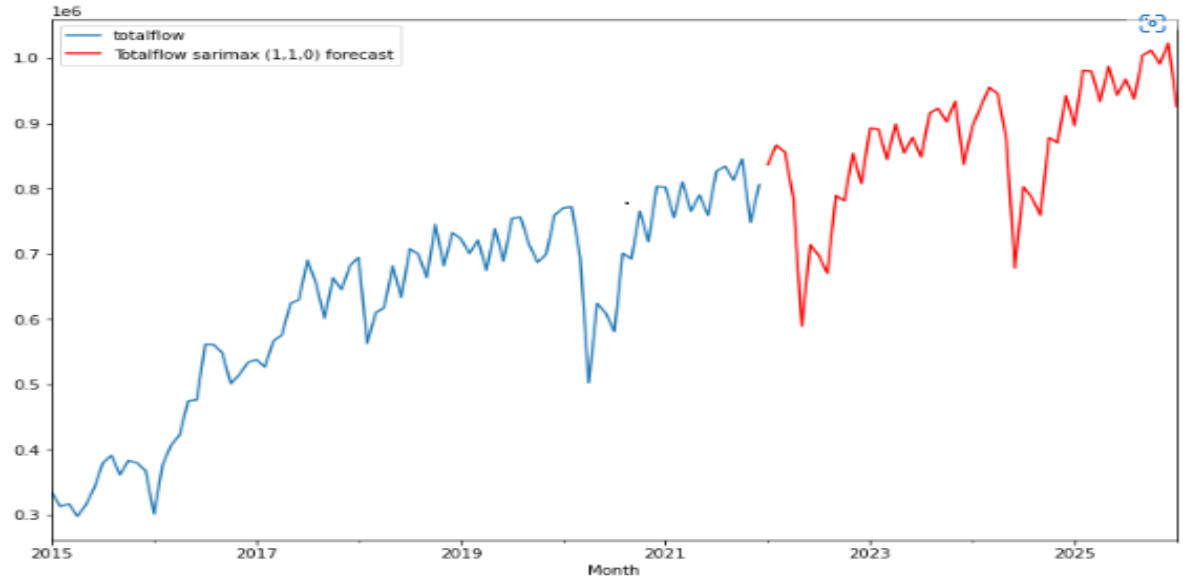


Figure 29 Graphical visualization of Actual monthly traffic flow and future forecast using SARIMA

5. Flow chart diagram of Toll collection system in ETRE

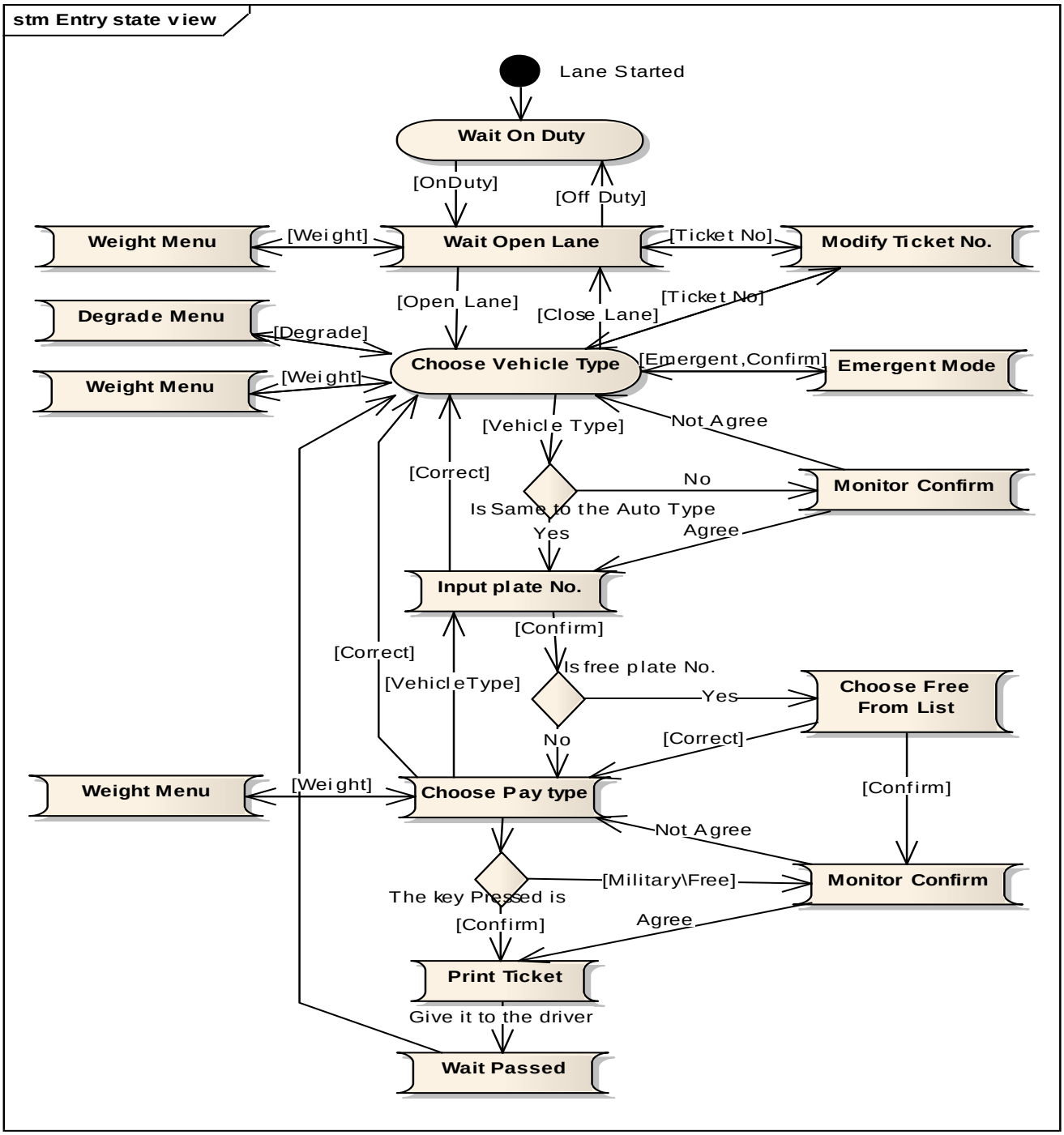


Figure 30 2.2ETRE Toll Collection and vehicle registration process on Lane status and keys relation schema

6. Tulu dumtu toll station LAN architecture

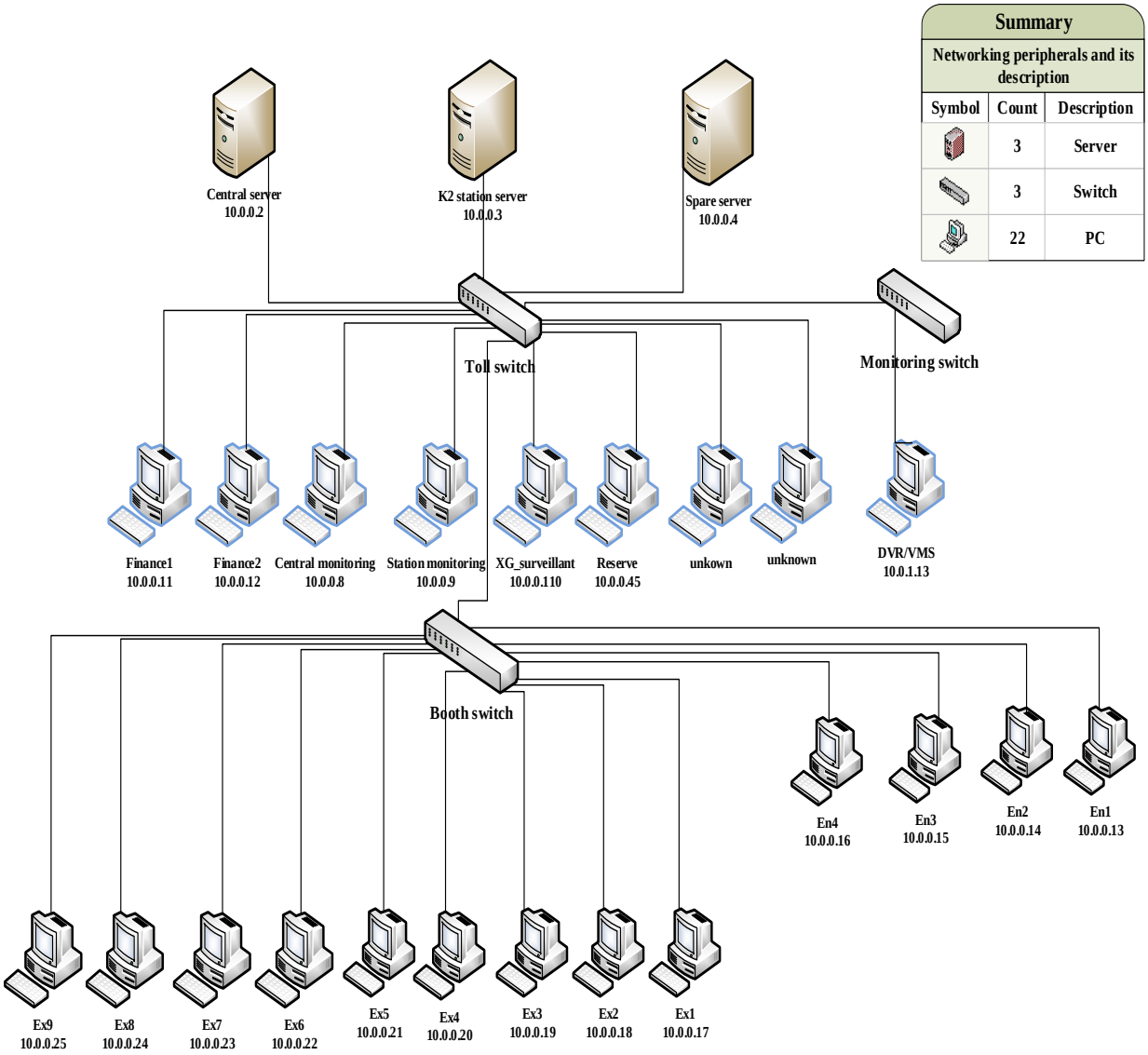


Figure 31 Tulu Dimtu Toll Station Lan Archtecture