Road Traffic Congestion Estimation using 3G Handover Data: for the case of Addis Ababa

By: Chalachew Berihun Teshale

A Thesis submitted to

School of Electrical and Computer Engineering

Addis Ababa Institute of Technology

In Partial Fulfillment of the Requirements for the Degree of Master of Science in

Telecommunication Engineering

February 23, 2020

Declaration

I, the undersigned, declare that the thesis comprises my work in compliance with internationally accepted practices; I have fully acknowledged and referred to all materials used in this thesis work.

ChalachewBerihunTeshale _____

Name Signature

Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

This is to certify that the thesis prepared by ***ChalachewBerihunTeshale***, entitled *Road Traffic Congestion Estimation using 3G Handover Data: for the case of Addis Ababa* and submitted in partial fulfillment of the requirements for the degree of Master of Science in Telecommunication Engineering complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Internal Examiner_____ Signature _____Date _____

External Examiner_____ Signature _____Date _____

Advisor MesfinKifle(PhD) Signature_____Date _____

Director of Post _____ Signature _____Date _____

Graduate Program

_____

Dean, School of Electrical and Computer Engineering

Acknowledgments

I would like to express sincere gratitude to my advisor for patient support, encouragement, and guidance throughout the year of the study.

I would also like to thank Dr. Dereje for providing me research area for study.

Equally, Thanks to all who helped me through the study, Especially, a special thank you goes to my family for the encouragement I get, which can not be priced.

Abstract

There is an increase in vehicle use in Addis Ababa which increases by 5% annually, not only this but also the urban arterial roads are of mixed type and there are no automated means that monitor the road traffic congestion. This creates road traffic congestion which is one of the main problems that affects everyone in time, energy and resource management. Road traffic estimation is instrumental in developing a more advanced transportation system to avoid congestions, trip delay, improper fuel usage, and vehicle use. There are different methods to collect data and estimate the degree of road traffic congestion such as manual counting (observation), global positioning system, inductive loops, using a camera system. The latter three mentioned are expensive in installation and maintenance with coverage limitations. This thesis focuses on one of the cost-effective alternative methods, which is based on 3G cellular handover data in the case of Addis Ababa's main selected road for the experiment.

In this study, we examined this alternative method to estimate the degree of road traffic congestion using a simple feedforward backpropagation neural network by considering and ignoring the altitude parameter. Related research has been done, but this study does not show by considering the altitude as a factor. We collect handover data while driving along selected arterial roads in Addis Ababa, and also handover archived data has been used for some sections of the streets for the experiment. For the congestion estimation data classification, AADT(Annual average daily traffic) data has been collected from AAPC(Addis Ababa police commission), and LOS(level of services) also used, which is based on the speed of test vehicles. The neural network was then trained and tested using the collected data against the road level classification done. The results found showed better performance of congestion estimation with an accuracy of 92.5% when the feature altitude is considered whereas without considering this variable, the result found is similar to those of other related research works 82.1 %. It has been shown from the effects that the improvement in the accuracy of estimation increases by 10%.

Keywords: Cellular networks, 3G handover, GIS, Weka, MLP

Contents

List of Figures

List of Tables

ABBREVIATIONS

2G	Second Generation
3G	Third Generation
AADT	Annual average daily traffic
AAPC	Addis Ababa Police commission
ANN	Artificial Neural Network
BPNN	Back Propagation Neural Network
CDR	Call Detail Record
CDT	Cell Dwell Time
CID	Cell ID
DT	Drive Test
FCD	Floating Car Data
FL	Fuzzy Logic
GIS	Geographical Information System
GPS	Geographical Position System
HO	Handover
HOD	Handover Data
HOFC	Handover Frequency Count
LAC	Location Area Code
LOS	Level of Service
LSTM	Long Short Term Memory
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MLP	Multilayer Perceptron
OSS	Operational Support System
PR	Precision-Recall
RNN	Recurrent Neural network
ROC	Receiver Operating Characteristic
SPSS	Statistical Package for the Social Sciences
UE	User equipment
UMTS	Universal Mobile Telecommunication System
UTRAN	UMTS Terrestrial Radio Access Network
WCDMA	Wide-band CDMA

Chapter 1 Introduction

This chapter will present the introduction part of congestion estimation, problem statement, objective, methodology, scope, and limitation of the research, with the contribution of the study and related literature review.

Traffic congestion is a condition on transport networks that occurs as use increases and is characterized by slower speeds, longer trip times, and increased vehicular queueing[1]. When traffic demand is high enough that the interaction between vehicles slows the rate of the traffic stream, this results in some congestion. As congestion increases, the hours spent on the road also increase. The cost of this congestion, measured in wasted time and fuel, incurs Addis Ababa annually about 5-8 Million Birr per intersection only for vehicle and fuel cost[2].

There are different methods to estimate the degree of road traffic congestion such as manual counting (observation), global positioning system, and inductive loops and using a camera. All mentioned in the above are expensive in installation, maintenance with coverage limitation, an alternate way is using cellular handover data specifically 3G by using machine learning techniques. The reason for using 3G handover data is due to its coverage dominance in the city of Addis Ababa.

The Universal Mobile Telecommunication System (UMTS) is a 3G mobile communications system that provides a range of *Broadband* services to the world of wireless and mobile communication users[3]. It is a widely adopted 3G wireless cellular standard. Wide-band CDMA (WCDMA) is the air interface for UMTS the general architecture of the mobile network is shown in Figure 1-1.

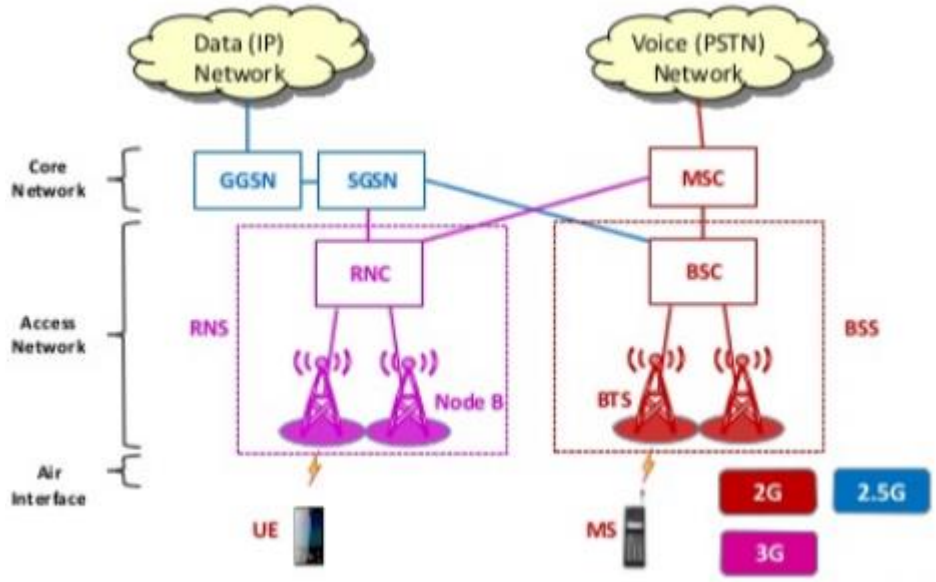

Figure 1-1. The Cellular Network Architecture

Some brief definitions of the acronyms related to the UMTS network are provided below:

Node B: The base transceiver station in the UMTS system is called Node B. Node B is the hardware that communicates directly with mobile handsets.

RNC: The Radio Network Controller, which is hardware that responses for controlling all node Bs that are connected to it.

Iu: Iu is the interface between the RNC and the Core Network.

Iub: Iub is the interface between the RNC and Node B.

The entire experimental log data set of this study was provided to collect handover logs on the UMTS on the air interface.

1.1 Congestion estimation

As mentioned early on, there are different types of congestion estimation methods. This section discusses the alternative means and cost-effective one which is by using cellular handover data. There are different categories of handover, the classification of handover can be based on signaling character, properties of the cell the same and different frequency use and purpose of handover (mobility of user equipment).

Mobile stations in a cell border that are transmitting at their maximum power cannot increase their power levels. Hence, they may get disconnected if they move far from the serving node further unless an HO takes place to other neighboring cells. To make a mobile station to be continuously connected to a cellular network when a mobile station user travels from one area of coverage or cell to another cell within call duration, HO should be initiated. This kind of HO can be related to mobility. Another condition that triggers HO is the malfunctioning or performance degradation of a serving cell. Due to this, mobile users who were previously connected to a current faulty cell are forced to connect to a nearby cell with a better signal level among other neighboring cells. Both mobility and malfunctioning cells can be because of HO occurrence which is due to received signal strength level drops below a particular threshold value set by an operator. The strength of receiving signal is influenced by fading due to shadowing, and destructive interference (reflection, refraction, and scattering at small obstacles). Fading is a variation of the attenuation of a signal with various variables. There are two main categories of HOs, which are classified as inter-cell and intra-cell HO.

The purpose of inter-cell HO is to keep the signal quality and coverage when the user moves to a new cell area whereas the use of intra-cell HOs is to change one channel inside the existing cell, due to fading or interference, to a new channel with better conditions. Basic types of UMTS handover are hard handover, soft handover, and softer handover. In this thesis, HO data is the primary input for the road traffic estimation. The next sections discuss those handover categories.

A. Hard Handover

As the name indicates, hard HO is a hard change of connection. Where one link is broken and another established. It means that all the old radio links in the UE are removed before the new radio links are found (break-before-make). Hard HO can be seamless or

non-seamless depend on its noticeability to the user call. If an HO requires a change in carrier frequency (i.e., inter-frequency HO), then it is always categorized as hard HO Figure 1-2 shows the hard handover scenario when break-before-make is applied when the user's equipment communicates with only just one Node B. Connection with the old Node B is broken before the new connection is established[4].

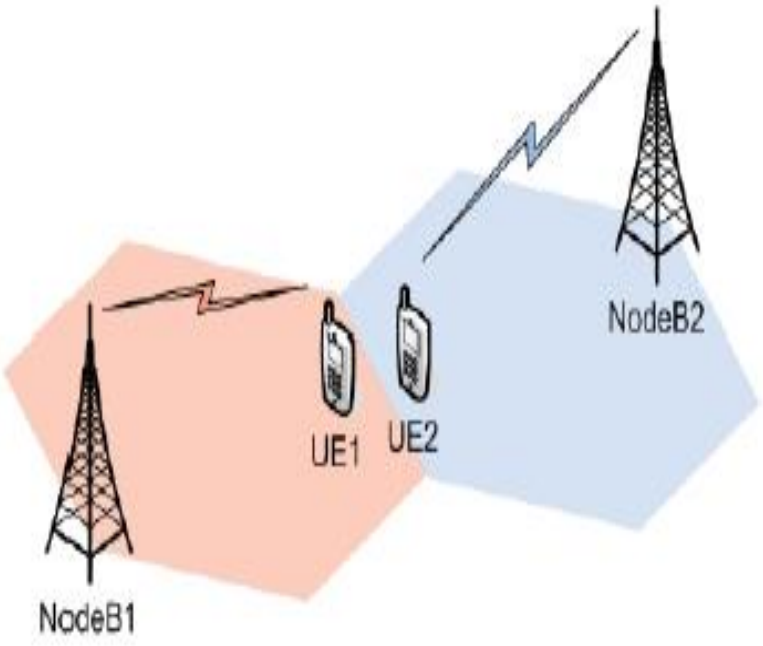

Figure 1-2. Hard Handover Scenario[4]

B. Soft Handover

Soft HO means that the radio links are added and removed in a way that the UE always keeps at least one radio link to the UTRAN (make-before-break). It occurs when a UE is in the overlapping coverage area of two cells. Soft HO is performed utilizing macro-diversity, in which several radio links are active at the same time. To implement Soft HO, it required that source and target cells need to operate on the same frequency. Figure 1-3showsSoft Handover scenarios. UE2 is located in the coverage area of two or more different Node Bs. The UE concurrently communicates with two or more Node Bs via two or more radio channels. A received signal in Node B is routed to the RNC and the RNC compares the signal on the frame by- frame basis[4]. The best frame is selected for the next processing; the other frames are discarded.

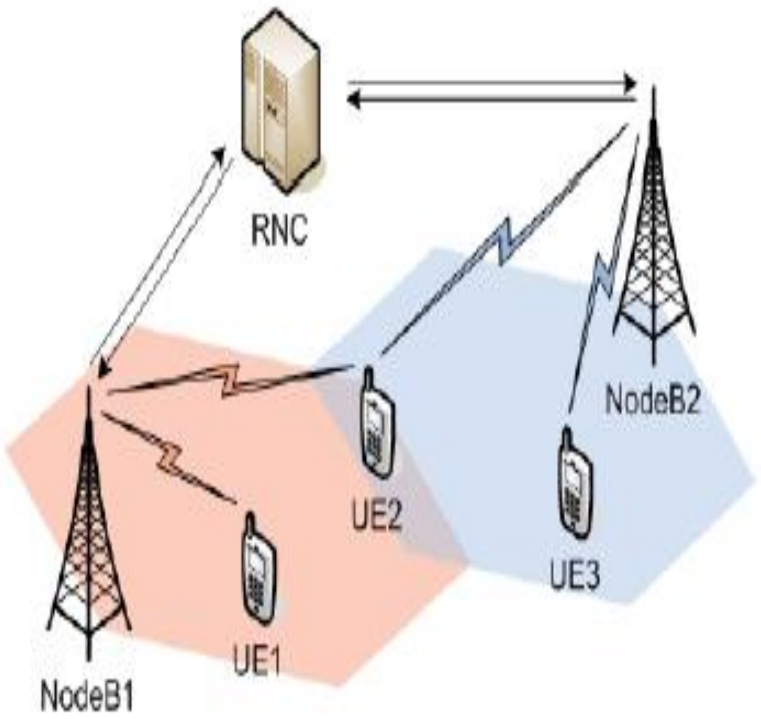

Figure 1-3. Soft Handover Scenario[4]

C. Softer Handover

Softer HO is exceptional cases of soft HO where the radio links that are added and removed belong to the same Node-B (i.e., this occurs when several sectors may be served from the same Node-B). During softer HO, a UE is in the overlapping cell coverage area of two adjacent segments of a base station, and it makes concurrent communication with two air interface channels (one for each sector that carries two separate codes in the downlink direction). In softer HO, macro-diversity with maximum ratio combining can be performed in the Node-B, whereas in soft HO on the downlink, macro-diversity with selection combining is applied[5][6]. Figure 1-4 shows how the softer handover will happen, softer handover is similar to soft handover. The main difference between these two handovers resides in the fact that a UE is located in the coverage area of two sectors of one Node B in the case of UE2, and Figure 1-5 shows handover in motion which can be any of the handover types mentioned in this section. This study focuses on those handover types, which are all horizontal ones.

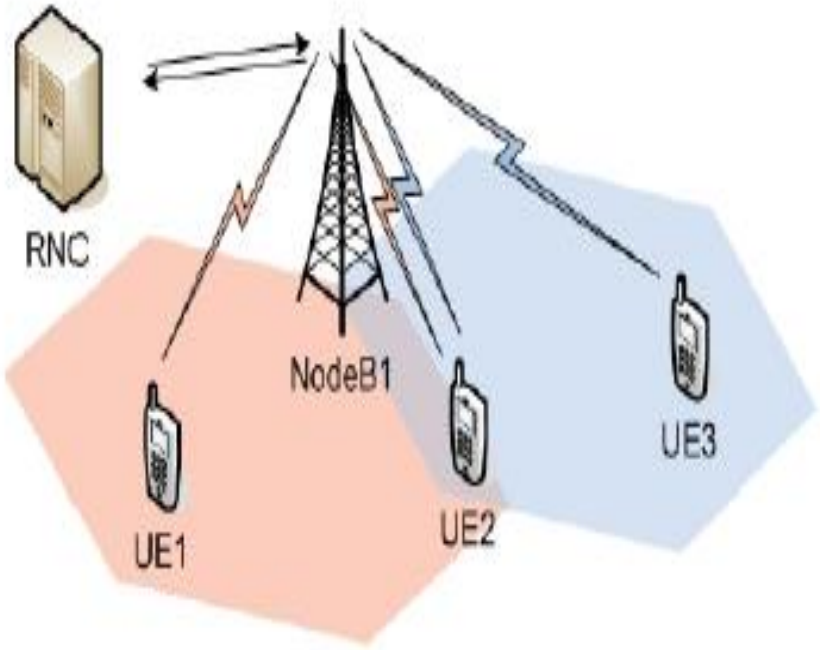

Figure 1-4. Softer Handover Scenario[4]

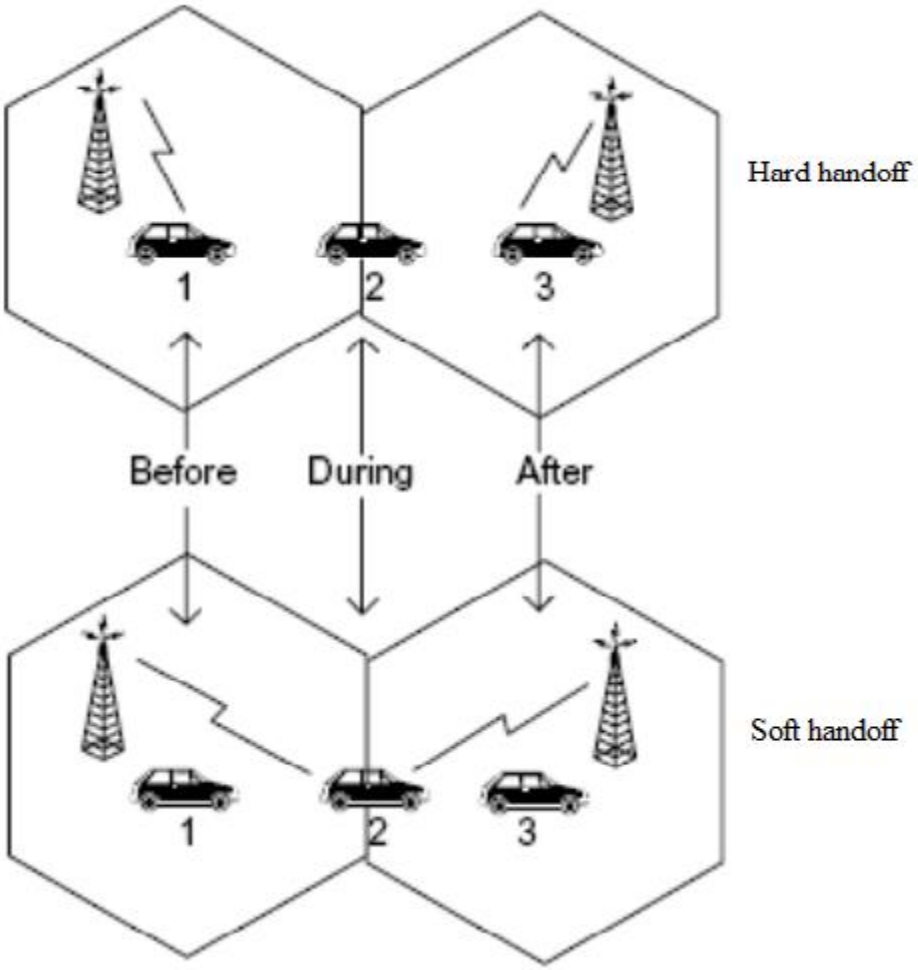

Figure 1-5. Handover in Motion[4]

In case of hard handoff, as illustrated on the above picture there are two node B's considered for the scenario, while the user is in motion with vehicle as shown the car is in motion at that time the serving node B indicated on cell area one, which means before hard handover starts, on the border cell of one and three hard had off will be performed and then after the serving of node B cell one will be migrated to the serving cell of node B's of cell area three. As indicated on the cell border area two where break before the connection will take place. In this case, the user notices that the service is interrupted.

UE is located in the coverage area of two or more different Node Bs. The UE simultaneously communicates with two or more Node Bs via two or more radio channels. A received signal in Node B is routed to the RNC (Radio Network Controller). The RNC compares the signal on the frame by- frame basis. The best frame is selected for the next processing; the others frames are discarded, what makes it different from the hard handoff is that break before the connection

will not happen rather active serving cell migrate or replaced by the strong signal of dedicated cell and this will happen without the user senses the service interruption.

1.2 Problem Statement

The total number of ***vehicles registered*** in the capital city of Addis Ababa passed half a million, with an increase of 5% annually[7]. Road Traffic congestion affects everyone in *time, energy, and resource management*. Research results show that, on average, four working hours per day wasted due to road blockage. In Addis Ababa, there is no automated way that can monitor the level of road congestion. The absence of a computerized road traffic monitoring system is one of the reasons for trip delay, traffic accidents, fuel, and improper vehicle utilization. Addis Ababa need to have well-organized technique so that the road users and property owners can minimize the level of congestion whether the congestion is due to accident, unbalanced infrastructure, blockage or special public gathering that face day to Day especially in peak hour, Table 1-1 shows the six-year statistical data of AAPC(Addis Ababa Police Comision) road traffic accidents due to different factors including congestion.

Table 1-1.The Six-Year Statistical Data of AAPC[8]

ID	Injuries	Property damage	Cost of traffic Accidents
1	14210 (2003-2005)	37142	Not Stated
2	18722 (2009-2016 G.C)	85316 (2009-2016 G.C)	1.87 billion Birr (2009/2010 E.C)

Other researchers in Portugal and Bangkok study on the use of Handover data of cellular network for road traffic congestion estimation using cell ID, handover frequency count or cell stay time and Location area code, their study does not consider altitude as one of the factors[9],[10]. The main goal of this research is to study the impact of altitude on the accuracy of road traffic congestion estimation.

1.3 Objective

General objective

- To study the impact of altitude on the estimation of road traffic congestion via 3G handover data using machine learning techniques.

To meet the general objective specific objectives are listed:

- Collect 3G handover data with sample urban arterial road
- Collect AADT data about the level of road traffic congestion from AAPC
- Extract the required attribute from the hand over data
- Select appropriate machine learning algorithms
- Study the impact of altitude on the road traffic estimation accuracy using Machine Learning technique

1.4 Methodology

The primary purpose of this research is to study the impact of altitude on the road traffic congestion estimation by developing a model using ANN by considering the recommended feature altitude. To achieve the objectives of the research and answers the research questions, the following method designed. Reviewing Academic literature on road traffic estimation. Preparing data set, which is collected from 17-kilometer distance coverage of 3G Handover data and relevant features selected for the research experiment. Then after the selected features were preprocessed to make it ready for the analysis. Finally, datasets were labeled and prepared for training and testing the ANN models. The developed models were tested and evaluated. Results were discussed and reported. In the end, outcomes were summarized, and recommendations were provided. Figure 1-6 shows the methodology adopted for the research work.

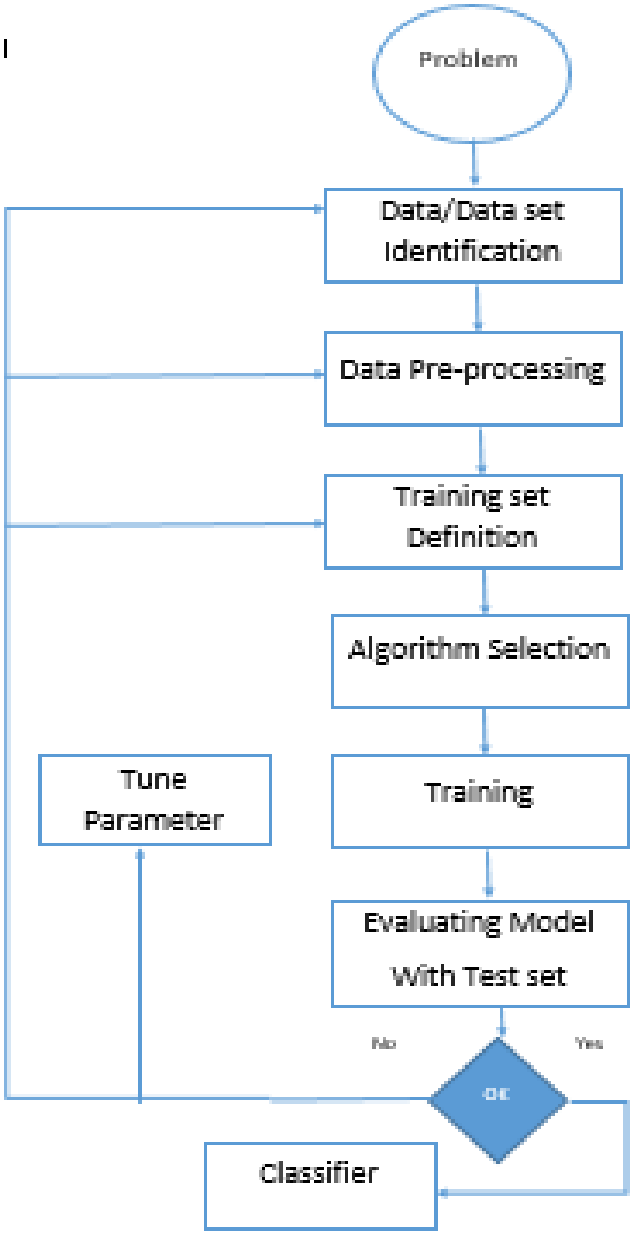

Figure 1-6. Research Methodology Adopted[11]

1.5 Scope and Limitation

There are many options for road traffic estimation data collection processes, such as inductive loop, GPS, camera systems, radar, and many more. However, this study focused on the estimation of road traffic congestion using 3G handover data by considering the impact of altitude. The congestion level of classification is in two ways which are moderately congested and congested. The third class of classification is not covered in this study which is free from

congestion due to data unavailability and the handover data considered is the horizontal type with one way of road data collection that can be a limitation.

1.6 Contribution

The result of the research will have practical contributions. The research shows the level of importance of considering altitude in road traffic estimation using the handover data. The estimation level is determined by the algorithm and techniques used in the research. This research could be used as a reference for farther study.

1.7 Related work

M. G. Demissie *et al.* presented a method for estimating hourly traffic volume through a combined use of average traffic volume and cellular networks handover information and tests the technique in such a way that, handover data were collected from cellular towers from the country of Portugal for 16 Km road distance, 487 cell towers used for three days with restricted pedestrian. The traffic sizes were also obtained from 101 traffic counters and then, applied statistical analysis with a regression model to investigate the relationship between vehicular and cellular traffic and the result found is MAPE (Mean Absolute percentage error) 18% that handover data can be used as to detect vehicle movement [9].

W. Hongsakham *et al.* proposed alternative methods for estimating degrees of road traffic congestion by using Cell Dwell Time (CDT), which is the information available from cellular networks. 12-hour data with 708 data sets of signaled road used. Measurements of Cell Dwell Time were taken and classified into one of the three degrees of congestion to users, which is "recommended" for use, "not recommended" condition, and "avoid" traffic condition. Furthermore, W. Hongsakham *et al* uses K- means clustering algorithm and backpropagation neural network. The paper indicates, both the K-means clustering algorithm and the backpropagation neural network provide promising estimation accuracy shows that with the neural networks approach producing good results 87.67 %. Traffic management agencies carry out different types of traffic counts to obtain quality information about traffic congestion and for effective traffic management. Active traffic management requires a costly means with monotonous operations. There are different methods, such as the use of manual counting, global positioning system, radar, inductive loops are all costly concerning time, personnel, and complexity. W. Hongsakham *et al* states that after handover, data were collected from cellular towers estimation of road congestion performed and the result concluded that the study

encourages the exploration of the use of cellular network handover information in estimating road traffic volume. Finally, the paper state that the neural network method works better[12].

W. Pattara-atikom *et al. show* an alternative method to estimate the degree of road traffic congestion based on a new measurement metric called Cell Dwell Time using a neural network. A series of CDTs measures were taken along roads in the Bangkok metropolitan area. The human judgment of traffic conditions was also made after training the neural network and then tested using the collected data against human perception. The results showed promising performance of congestion estimation with an accuracy of 79.43%. The experimental results show that CDT data have the potential to be a traffic congestion estimation measure. The author mentioned that the main advantage of using CDT is low implementation cost as no additional expensive hardware infrastructure is needed. Besides, geographic penetration of mobile phones and their related infrastructure is virtually 100% on the test road where the experiment took place[13].

The handover data of the cellular network can be used as important data for urban traffic management since most of the urban traffic users are mobile users this opportunity can be used as a count of traffic users on arterial roads, using the handover data the author developed a system that can classify users of cellular network in different categories , the developed algorithm classifies mobile phone users like pedestrians, slow-moving vehicle users, daily travelers, users of large gathering and non-resident users[14].

F. M. S. C.P.IJ. van Hinsbergen *et al.* starts, how it is hard to select the most appropriate road traffic estimation method for one particular purpose. The objective was to provide a taxonomy of all different approaches stated in different literature as a practical reference for research purposes. The paper mention the prediction models in three categories which are: the first is Naïve(clustering) method which is without any model assumption and the one which is easy and low accuracy, the second method listed is Parametric(traffic models) Model is predetermined, and the last is Non-parametric (like ANN) this method estimate with flexible parameters. And finally, the author concluded that apart from traffic simulation models only a few methods are used for network-wide predictions, while this is necessary for most practical applications. In addition to this, no single best method available, it depends on the data source type[15].

B. Sharma *et al.* developed a short term traffic forecasting model using a backpropagation artificial neural network for two lanes undivided highway with mixed traffic conditions in one

of a developing country which is India. The results found were compared with random forest (RF), support vector machine (SVM), the k-nearest neighbor classifier (KNN), regression tree and multiple regression models. Sharma et al. found that the back-propagation neural network performs better than other approaches used and achieved an R square value of 0.9962, and conclude as a good score[16].

Moreover, this research tries to address for estimating the road traffic state using 3G cellular network handover data by considering as well as disregarding altitude as a factor in the case of Addis Ababa. Figure 1-7 shows a summary of the related work review, as indicated in the first column are data sources for the purpose of traffic estimation followed by different prediction models used and the values considered.

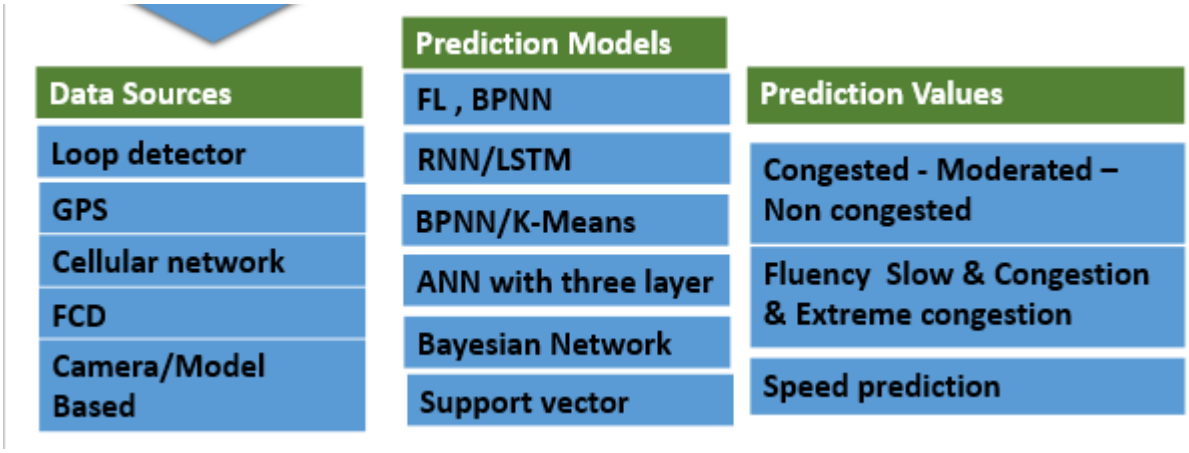

Figure 1-7. Summary of Related Work

1.8 Thesis Organization

This thesis contains six chapters. The second Chapter, road traffic and traffic information system includes road classification and capacity in general. Chapter 3 Material and method deal with the concepts of machine learning, machine learning methods, which are used in this study. This chapter also discusses Data preparation, which deals in collecting describing preprocessing the data to make it ready for experimentation. Chapter 4, the Modeling and experimentation focus on experimentation, which includes building models, testing and evaluating the models. Chapter 5 discuss the outcomes of the experiment followed by result and discussion. The last chapter is dedicated to conveying conclusions and recommendations of the study as future work.

Chapter 2 Road Traffic and Traffic Information System

In the previous chapter, we have discussed the introduction part of congestion estimation, problem statement, objective, methodology with the contribution of the study and related literature review. This chapter presents road traffic and traffic information system overview to compare with the data collection method used in the study.

Congestion, become common characteristics in the urban road transportation system of Cities in developing countries which result in high operating cost, loss of users productive Time, and more fuel consumption, among others[17].

First: the causes of car traffic congestion are several and interconnected factors, like abrupt urbanization, that concentrate individuals and economic activities in urban areas or Cities.

Second, as a result of different dispersed, however, interconnected land-use patterns or Specializations of the urban areas in some activities, for illustration, the labor force intense in some areas, residential and recreational areas also in another faraway place that makes people move between them.

Third; a mismatch between supply and demand; the problem is severed in peak hours in specific as most people start and end their work at the same time in the mornings and evenings.

Due to various factors, road congestion is becoming a more severe problem in the capital of Ethiopia from time to time. It is clear that these added portions of the society also need transport services to attain their day to day activities. Nevertheless, the city is incapable of handling the existing high transport services demand. Besides, inefficient land use planning, poor infrastructure, and the absence of sound traffic management are the primary reasons for the problem of traffic congestion. As a result, realize the present situation of vehicle traffic congestion is a critical area of consideration to make the right decision to solve the issue and thereby sustain unified traffic flow to contribute to the economic growth of the city is an urgent issue.

2.1 General Road Classifications

There are several primary road classifications in urban areas, and they are based on:

1. Traffic: such as volume

2. Function: functional importance in the whole networks

3. Administrative: on the jurisdiction of individual administrations on various roads.

2.2 Road Capacity

The road capacity can be identified by using the level of service and annual average daily traffic. The next section discusses these two methods.

A. Level of service

The magnitude of traffic congestion is differing from one level-of-service to another level-of-service. The road becomes more congested and incapable if the volume of vehicles increases from “A” level- of- service to “F” level- of- service. A is an ideal level, and F is the worst level of service. It is used to investigate roads and intersections by categorizing traffic movement and assigning quality levels of traffic based on performance measures like vehicle speed, density, congestion, etc. Table shows LOS with respective description[17][18].

Table 2-1.Los Description Table[17][18]

ID	LOS	Description
1	LOS A	Density is low enough, Individual vehicles flow freely, not affected by others.
2	LOS B	A stable flow condition or there is relative freedom in speeds.
3	LOS C	There is a relatively stable flow, but individual vehicles influence the Flow immediately and become significantly affected by interactions with other cars in the traffic stream.
4	LOS D	It is a crowded roadway situation as mobility and a stable flow is restricting with a large number of vehicles. Speed and freedom to movement are harshly restricted.
5	LOS E	No usable gap between vehicles, speed is slow, the condition can easily cross over into LOS F region Roadway accommodates nearly to its full capacity, low rate, and small increment in the traffic volume will affect the traffic movement more.
6	LOS F	Breakdown condition, number of vehicles arriving > number of cars leaving, speed is zero Vehicles move in a locked each other within the front and beyond Condition. Speed is mostly to zero, and the travel time cannot be predicted

The level of service for the case of Addis Ababa falls under the category of D and E at the time of peak hour for the selected sample road area, and this can be known using the DT data which comprises the speed as one parameter and speeds with less than 13Kmph can be considered congested based on the urban arterial road target speed.

B. Annual Average Daily Traffic

Annual average daily traffic (AADT) is a measure used primarily in transportation planning and engineering. It is the whole volume of vehicle traffic for a year divided by three hundred sixty-five days. AADT measures how busy the road is and is considered as one of the most crucial raw traffic datasets where it provides essential inputs for traffic model developments that can be used for congestion management.

There are two ways for AADT calculations based on the data count collected:

- Permanent automatic traffic recording stations deliver continuous counting of the traffic on selected roads (mostly on highways) for the entire year. The advantage is to provide traffic counts that are typically recorded in fifteen minutes or hourly intervals, seven days a week, and three hundred sixty-five days a year intervals. Accordingly permits an additional exceptional level of analysis and a more accurate annual average than short-term counts. A permanent automatic traffic recorder is the only way to provide exact AADT values (when used under perfect conditions).
- Short-term traffic counts (also known as seasonal, portable, or coverage counts) offer roadway segment-specific traffic count information on a cyclical basis for a large number of road segments. The gathering data period typically ranges from 1 to 7 days, where data are recorded in 15 min or hourly intervals. Because of differences in day-to-day variation in the traffic flow, the count period is reliant on the road on which it is situated, e.g., rural or urban. For variation minimizations, the minimum requirements could be fixed, for instance, at 48-hours of continuous data for rural counts and 24-hours of continuous data for urban counts[19]. The road traffic classification is done based on both LOS and DT for the experiment. Table 2-2 shows the sample road area with AADT, and Figure 2-1 shows the respective graph.

Table 2-2. Sample AADT

ID	Place name	AADT
1	Megenagna square	49274
2	22 Mazoria	25041
3	Meskel square	55526
4	Bambis	43992
5	Estifanose church	50392
6	Bole Medhanialem	38285
7	Imperial hotel	21841
8	Lem hotel	36634
9	Olympia	43992

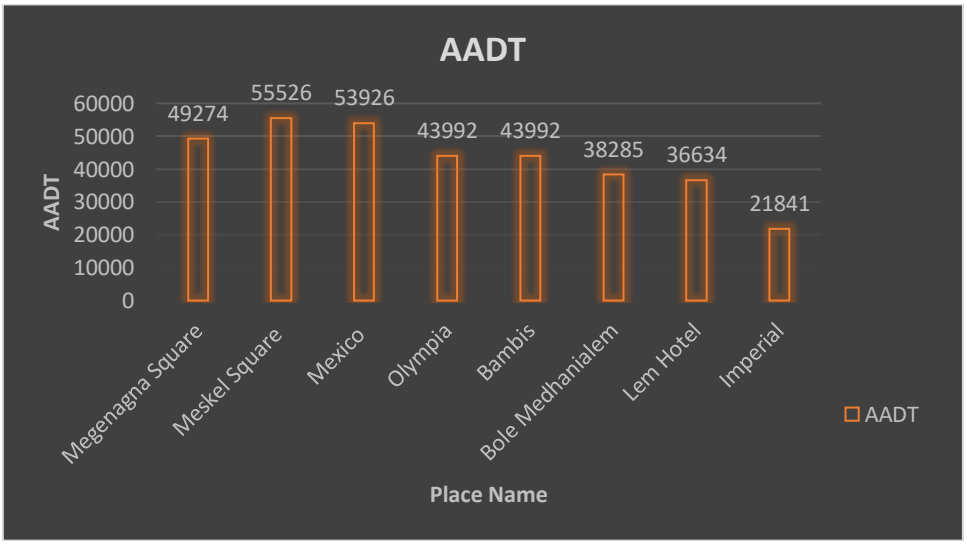

Figure 2-1. Graph Showing Sample AADT

This chapter tries to address the main root causes of road traffic congestion, primary road classification and how to identify the capacity level of different types of roads.

2.3 Road traffic information system Sensor Types

In this subsection, some common traffic information systems will be discussed with the respective purpose and limitations.

2.3.1 Loop detectors

Inductive loop detectors are built into the roadway so that they can detect each vehicle that passes over them. They work by detecting the metal of a vehicle as it passes over the detector. Properly calibrated, a loop detector is capable of providing high-accuracy flow and occupancy data, the latter of which can be used to infer density. When two loop detectors are placed close together, velocity can be measured by looking at consecutive crossing times. While the quality of the measurements from loop detectors is often good, filtering is still required from producing quality input data to highway estimation models. Loop detectors are not capable of directly measuring travel times. Loop detectors connected to an internet connection that can be used to transmit the data to a central server in real-time (that can subsequently be used in traffic information systems). Many locations throughout the developed country also have loop detectors placed on arterial roads. However, for arterial roads, it is very rare for the loop detector to be connected to the internet for easy transmission of the data to a central server. For this reason, arterial traffic information systems cannot rely on loop detector data as there is not enough of it to estimate conditions on the whole arterial network[20].

2.3.2 License Plate Readers

License plate readers are capable of spontaneously extracting the numbers and letters from momentary vehicles when placed right above a lane of traffic when multiple readers are set up at two points along the road, it is promising to extract travel time information for vehicles passing both locations. License plate readers suffer from the logistical problem of finding good locations to place them. When properly positioned and calibrated, these devices are capable of providing high-accuracy of travel times[20]. Due to the difficulty in placing these devices, they are not common throughout the roadway. Mostly they remain a data collection tool for specific studies.

2.3.3 Using GPS

Global Positioning System tracking (GPS) is a method of working out exactly where something is. A GPS tracking system could be placed in a vehicle, on the terminal of a cell phone. GPS provides information on the exact location of the device used as well as can track the movement

of the cars used. A GPS tracking system uses the Global Navigation Satellite System (GNSS) network. Therefore, a GPS tracking system can possibly give both real-time and historic navigation data on any kind of journey[21].

2.3.4 Video Camera detection

Traffic-flow capacity using video cameras is an alternative form of vehicle detection. This type of system is known as a "non-intrusive" method of traffic detection[22]. The cameras are characteristically attached to poles close to the roadway. Most video detection systems need initial configuration to "teach" the processor the baseline background image. This usually includes entering well-known measurements such as the altitude of the camera above the roadway. The distinctive output from a video detection system is lane-by-lane vehicle counts, speeds and lane use evaluations.

Loop detectors need to be connected to an internet connection and it is very rare to be connected always. Whereas the license plate readers suffer from finding a good location to place them. Overall alternative data collection methods need installation and maintenance. As mentioned early, cellular handover data for traffic collection and estimation is cost-effective in installation and maintenance compared with other related traffic information data gathering methods.

Chapter 3 Materials and Methods

In chapter 2 we have been presented road classification and road capacity, this chapter discussed machine learning, materials, and methods.

3.1 Machine Learning

Machine learning (ML) is a group of algorithms that permits software applications to become more precise in estimating and predicting results without being openly programmed. The simple idea of machine learning is to shape algorithms that can take input data and use statistical analysis to estimate an output while informing outputs as new data becomes accessible. Learning the road congestion level or classification is vital for road congestion estimation.

This upcoming section discusses the introductory concepts, approaches, and techniques of algorithms applied in this thesis work.

3.2 Types of Machine Learning

Different machine learning methods can be supervised, unsupervised, semi-supervised, or reinforcement. This section primarily focused on supervised machine learning techniques, which are the experiment performed, but discuses other mentioned learning types also.

3.2.1 Supervised Learning

Supervised models can be described as learning a function as shown in **Error! eference source not found.**

.

$$f(x) = y \qquad \text{(Equation 3-1)}$$

Where y is the class of the data, and x denotes the features. Supervised learning models are trained with pre-classified data. The examples of input or output functionality are referred to as the training data. The supervised learning methods are categorized based on the structures and objective functions of learning algorithms, and ANN is on this category. In the upcoming sections discussed ANN. The supervised learning algorithms are trained using labeled examples, such as an input where the desired output is known. For example, a piece of equipment could have data points marked either “congested” (C) or “moderate congested” (MC). The learning procedure or algorithm receives a set of inputs along with the comparable exact outputs, and the algorithm learns by comparing its actual output with the right outputs to find errors, it then modifies the model accordingly. Methods like classification, regression, prediction, and gradient boosting, supervised learning uses patterns to predict the values of the label on additional unlabeled data[23].

3.2.2 Unsupervised learning

Unsupervised learning is used in contradiction of data that has no historical labels. The method does not now the right response. The procedure must figure out what is actually shown. The aim is to explore the data and discover some arrangements inside. Unsupervised learning works well on transactional data. Popular techniques include self-organizing maps, nearest-neighbor mapping, k-means clustering, and these algorithms are also used to segment text topics, recommend items, and identify data outliers.[23].

3.2.3 Semi-supervised learning

This learning method used for the same applications as supervised learning. Nevertheless, it uses both categorized and unlabeled data for training typically a small amount of labeled or categorized data with a large amount of unlabeled data. This type of learning used with methods such as classification, regression, and prediction, and useful when the cost associated with labeling is too high to allow for a fully labeled training process, application examples of this include identifying a person's face on a webcam[23].

3.2.4 Reinforcement learning

Reinforcement learning used for different applications like gaming, automation, and navigation. With this learning method, the algorithm learns through trial and error, which actions yield the most significant rewards. Reinforcement learning has three main modules: the agent which is the learner, the environment, and actions. The objective is for the agent to choose

activities that maximize the expected prize over a specified amount of interval of time. The agent will reach the goal much faster by following a good policy, consequently, the goal in reinforcement learning is to learn the best system[23].

3.3 Artificial Neural Network

ANN are the principal tools used in ML, and they are brain-inspired systems intended to replicate the way humans learn. Neural networks represent a brain symbol for information processing[24]. ANN learns through an iterative process of adjustments applied to its synaptic weight and bias level. They are also able to improve their performance through learning. ANN learning paradigm is either supervised or unsupervised. In the case of supervised, there is a need to train or teach the input and output pattern.

When ANN is used as a supervised machine-learning method, efforts are made to determine a set of weights to minimize the classification error. The objective of ANN is to reduce the mistakes between the ground-truth Y and the expected output f(X; W) of the network, as indicated in **Error! Reference source not found.**.

$$E(x) = (f(x; w) - Y)^2 \qquad \text{(Equation 3-2)}$$

The performance of ANN depends on both the weights and the transfer function.

3.3.1 Perceptron

Perceptron is the most basic form of a neural network, which consists of a single neuron that can receive multiple inputs and produces a single output. Perceptrons are used to classify linearly separable classes. As illustrated in Figure 3-1, a perceptron takes a vector of real-valued inputs, calculates a linear combination of these inputs, then outputs 1 if the result is more significant than some threshold and -1 otherwise using the selected function. The exact learning problem is to define a weight vector that grounds the perceptron to yield the accurate output for each of the given training examples[25]. The typical approach that the perceptron algorithm used for learning from a group of training instances is to run the algorithm recurrently through

the training set till it finds an estimation vector that is correct on all of the training sets. This prediction rule then used for predicting the labels on the test set of the data.

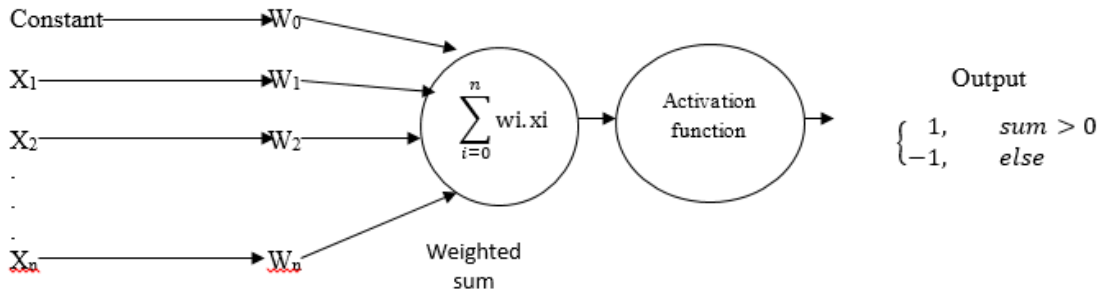

Figure 3-1. A Perceptron With Multiple Inputs And Single Output[26]

In the perceptron model, the weighted sum is calculated using, as shown in Equation 3-1 below, then evaluated and passed to an activation function, which compares it to a predetermined threshold theta. If the weighted sum is higher than the threshold theta, then the perceptron fires and outputs 1, otherwise it outputs 0 (-1).

$$\sum_{j=0}^{n} wj.xj = w1.x1 + \cdots + wn.xn \qquad \text{(Equation 3-1)}$$

There are Varieties of activation functions that can be used with the perceptron, but the step, linear, and sigmoid functions are the most common ones as shown in Table 3-1.

Table 3-1. Activation function with respective equation

ID	Activation function	Equation
1	sigmoid	$f(x) = \frac{1}{1+e^{-x}}$
2	linear	$f(x) = x$

All mentioned above, activation functions are triggered at a threshold $\theta = 0$. However, it is more convenient to have a limit other than zero. For that, a bias b is added to the perceptron into the inputs. The role of this bias b is to move the threshold function to the left or right, to change the activation threshold. Training the perceptron aims at determining the optimal weights and bias value at which the perceptron fires.

3.3.2 Multilayer Perceptron

For Linearly separable class's problem type, a single perceptron can be used if the problem is not such a type; there will not be a solution. An issue of such kind can be solved using MLP[27]. ANN are famous due to their robustness, fault tolerance, ability to learn and generalize adaptability, and parallel data processing. Those enable them to solve complex non-linear and multi input-output relationship problems. Furthermore, they are useful in practical applications due to their ability to do non-linear mapping, parallel processing methodology, and ability to learn from the environment and their subsequent adaptability to the environment. When compared to other methods, ANN is superior as a modeling technique for data sets with non-linear relationships. There can be some issues noticed in ANN; some of them are having many local minima, and also finding how many neurons might be needed for a task is another issue that determines whether optimality of that ANN is reached. ANN design includes the specification of the number of hidden layers and the number of units in these layers. As good a starting point is to use a hidden layer, with the number of units equal to half the sum of the number of input and output units. The problem defines the amount of input and output units.

The ANN-MLP is composed of three layers, the input layer, the hidden layer, and the output layer as shown in

Figure 3-2 the input layer consists of input nodes that represent the system's variable. The hidden layer consists of nodes that facilitate the flow of information from the input to the output layers. Weight factors associated with each connector controls the tide. The output layer comprises nodes that represent the system's classification result. The values of the output nodes are compared with Limits to determine the output and classify each case.

Input layer Hidden layer Output layer

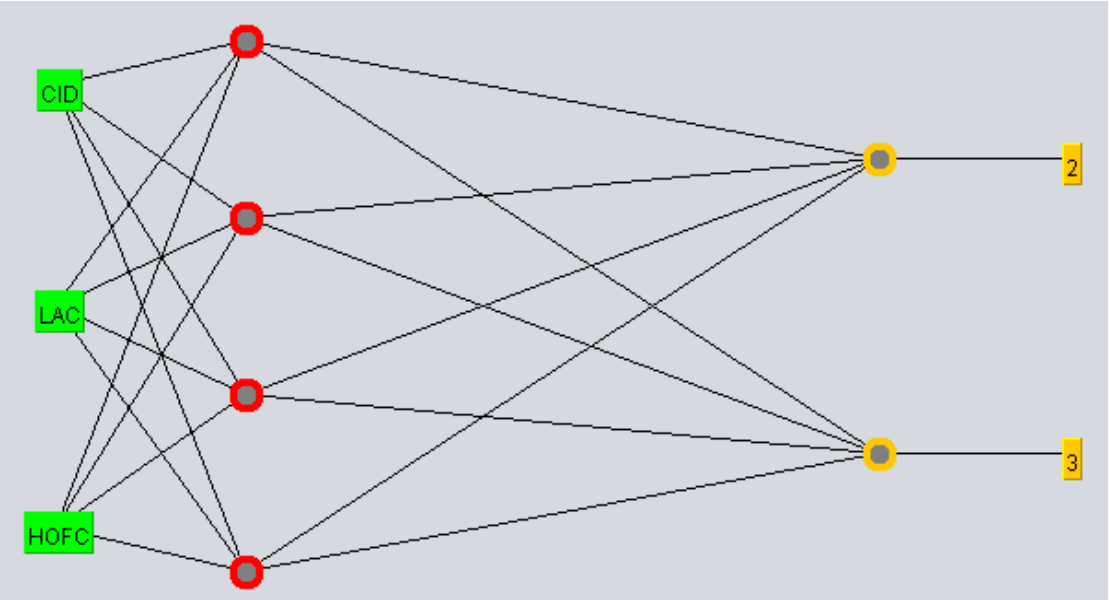

Figure 3-2. A Multilayer Perceptron Neural Networks

The weight adjustment, training process consists of running input values over the network with predefined classification output nodes. This process runs until the weight values are minimized to an error function or cost function. Testing samples are used to verify the performance of the trained network. Training is defined as the process of iterating through the training set to adjust the weights. For neural network training, random weights and biases are generated at first. Then, a training instance is passed to the neural network, where the output of each layer is given to the next layer until computing the predicted output at the output layer, according to the initial weights. The error at the output layer is calculated as the difference between the actual and predicted outputs. According to the error, the weights between the output layer and the hidden layers are corrected, and then the weights between the hidden layer and the input layer are adjusted in a backward flow. Another training instance is passed to the neural network and to the process of evaluating the error at the output layer, thereby correcting the weights between the different layers from the output layer to the input layer. Repeating this process for as many epochs will help in learning the neural network for the required accuracy of estimation.

Earlier to modeling, evaluation, and analysis, the raw data should be organized and has to be well-known for the experiment. For this research, Handover data of 3G were used, as stated

early, which is collected for three days totally for 10 hours and prepared for the experiment. One of the specific tasks in the classification problem model is getting an illustrative feature for the estimation of the road traffic estimation problem.

If there is too much unimportant and redundant data present or the data is noisy and undependable, then learning during the training part became more challenging. Data Preparation involves processing the data so that machine learning algorithms can produce a structural description of the information that is implicit in the data. It defines, process, and makes it suitable for the technique used. It is the first important step in data mining and plays a decisive role in the entire process[28].
The following sections discussed the data processing collecting, understanding, and preparing the data for model building the ANN-MLP.

3.4 Data Collection

The data collection process was done in four ways which are: first to collect from the operational support system for the weekends, and the second attempt was made to obtain from the different interface which is from the core network side and found to be incomplete. The third option carried out is using collected archived 3G handover data for possible selected sample road. The last option was collecting data by performing a driving test using the test terminal has also been done.

The handover data set used in the study was collected from Meskel square to Megenegan and SheroMeda Node B towers of ethio telecom that covers 17 Km long, totally including to the direction of Bole main road, sample roads are selected based on high congestion level, as shown in Figure 3-3.

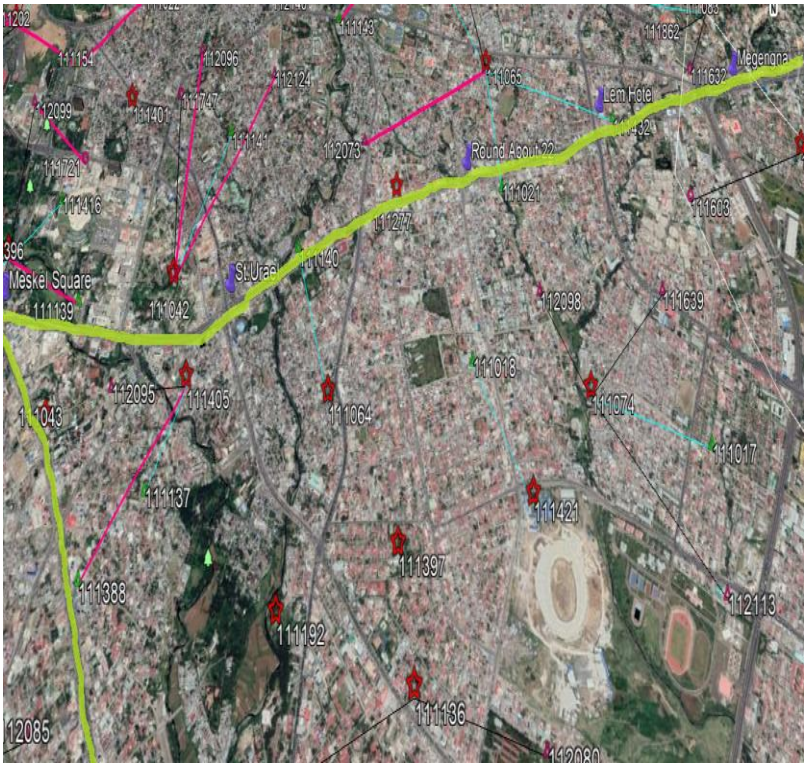

Figure 3-3. Sample Road Test

Data were collected in different ways as mentioned in this section one is using drive test terminal with GPS option enabled so that vehicle speed information can be captured which can be used for the classification purpose additional to AADT and the test is done during the peak hour which is on the range of 7-10 AM and 3S-6 PM shown in Table 3-2:

Table 3-2. Handover data used for the research

ID	Handover data used	Duration of test Data
1	June 2018	for 2 hour
2	March 2019	for 6 hour
3	July 2019	for 2 hours

And also additionally archived drive test data have been used for the experiment.

An example of data collected on the Samsung terminal is shown in Figure 3-4. That contains the altitude and speed of the test vehicle as a sample.

18Jun21 095219_Caller_MOS_Area_1.1:Qualcomm - Caller

Time	Latitude	Longitude	ServingCellRadioID	Altitude	SpeedKph	Uu_TimeBetween...
20:54:48:500	9.04648	38.72925		2501.000000	15	1040
20:54:49:500	9.04650	38.72919		2501.000000	18	
20:54:50:500	9.04652	38.72916		2501.000000	22	
20:54:51:500	9.04653	38.72912		2501.000000	22	
20:54:52:500	9.04656	38.72905		2501.000000	24	
20:54:53:500	9.04658	38.72901		2501.000000	22	
20:54:54:500	9.04660	38.72896		2501.000000	24	6520
20:54:55:500	9.04663	38.72890		2501.000000	26	
20:54:56:500	9.04666	38.72884		2501.000000	22	
20:54:57:500	9.04669	38.72879		2501.000000	18	3040
20:54:58:500	9.04671	38.72875		2501.000000	10	880
20:54:59:500	9.04672	38.72873		2501.000000	7	320
20:55:00:500	9.04674	38.72871		2501.000000	8	
20:55:01:500	9.04675	38.72869		2502.000000	9	
20:55:02:500	9.04677	38.72867		2502.000000	10	3921
20:55:03:500	9.04678	38.72865		2502.000000	7	

Figure 3-4. Sample Data Collected With Test Terminal

After collecting, the data extracted using Actix software so that to convert to the required format of CSV for further experiment.

The second way of data set used is from archived drive test (DT) data specifically on the area of the Sheromeda and Bole area. Additional to those data collections, the AADT data is collected form AAPC for road classification input.

The data collection process in the driving test is done with a continuous call using the terminal device of the road starting from Meskel square to Ayat, and the data is extracted and converted to the format required for the machine learning algorithm.

It contains the fields necessary for road traffic congestion estimation for data preprocessing to be successful features of cell id, date and time, location area code, active serving cell, altitude, longitude, latitude.

3.5 Data Selection

Data selection is a process, which requires understanding to choose useful features that capture the variability and essentiality of the data for the target ML algorithm to learn patterns from the data successfully. Besides, it has a vital role in reducing the complexity of the learning process.

As discussed, the collected HOD has fields with empty and duplicate contents, and some others are irrelevant for the intended thesis work. From the considerd fields, henceforward, the remaining four fields that are selected the most important for the study based the result found on the attribute selection result. Those selected fields are described in Table 3-3

along with the attribute selection result in Table 3-4 . Site ID is not considred because of CID comprehensiveness. parameters longitude and latitude are not considred for the study since those point have low correlation value.

Table 3-3. Description of Selected Fields

ID	Field Name	Description
1	CID	Cell Identification
2	LAC	Location area code
3	CST	Cell serving time (time between update)
4	Altitude	The geographical location of an active node

Table 3-4. Rank correlation value

Previously work	With new feature	Rank-correlation
CID	CID	0.9860
HOFC	CST	0.68714
LAC	LAC	0.97471
Time	Altitude	0.006391
	Time	0.000289

3.6 Data Preprocessing

Date preparation is responsible for identifying quality data. One of the critical stages in making ready the raw data to ML is data preprocessing, which might require data error elimination, data set feature selection, transforming the data set to the required format, and data set classification. The next sections discuss this process in detail.

3.6.1 Data Error Elimination

Data cleaning is to fill or remove the empty value of the data, eliminate the noise data, and correct inconsistencies in the data. There will be errors in the collected data set, and we need to address data quality issues. The collected data for this thesis, have some errors, such as incomplete values and 304 data set been cleaned.

3.6.2 Data Feature Selection

Theoretically, a dataset with more attributes in the learning process gives better results. However, in practice, this may not always be the case. Most machine learning algorithms are designed to learn appropriate features to use for making their decisions. Adding distracting

features confuses machine learning systems. In practical situations, there are many attributes for the learning process, some of them feasibly significant, and some are irrelevant or redundant. The problem is identifying a representative set of features from which to construct a classification model. For that reason, the dataset must be preprocessed to select useful attributes. Even though many learning schemes can select features appropriately and ignore irrelevant ones, but in practice, their performance might be affected. Because of the negative effect of irrelevant attributes on most ML algorithms, it is common to precede learning with an attribute selection. More importantly, dimensionality reduction yields a more compact, easily interpretable representation of the target concept, focusing attention on the most relevant features[29].

Feature selection is a process of selecting a subset of features from which to build an estimating model. The feature selection is usually made for model simplification, reducing training times and computational cost, and to help reduce the risk of over-fitting, and thus improve model generalization.

The approaches followed in feature selection for the study are filter method; it makes an independent assessment based on general characteristics of the data, attributes filtered to produce the most promising subset. Attribute evaluator with correlation-based to evaluate feature subset using the machine learning algorithm that will be employed for learning.

The main principle for feature selection with the filter method is that features with high correlation value are selected than that of the low value.

For this study, two subsets of features since the main objective of the study is to see the impact of altitude feature; the experiment is done with considering and disregarding it. In this preprocessing stage, we have tested those feature subsets. Furthermore, those the remaining feature that have better performance selected as study features based on the correlation value. Subsets of futures are listed in Table 3-5.

Table 3-5. Subsets of Feature Lists

ID	Other researchers feature	Features considered	Study Features considered
1	CID	CID	CID
2	HOFC	CST	CST
3	LAC	LAC	LAC
4	Time	Altitude	Altitude
5	CDT	Time	Time

3.6.3 Dataset Formatting

Formatting is a re-engineering process that arranges the input dataset into a format that is acceptable by the particular ML algorithm which is ANN or MLP. Attributes are of nominal or numeric data type, attaining format consistency in all of the records in the entire file is a critical issue. In our case, all the records used are of numeric format except the nominal class level, and consistency of format is checked carefully.

A data set classification generally comprises separating data into training and testing sets. Similarly, in this thesis, the preprocessed dataset was partitioned into two parts, training and testing, as indicated in Table 3-6. About 80% of the dataset instances are used for training and the remaining 20 % for testing.

Table 3-6. Data Set Partition

ID	Subsets Criteria	Training data set	Testing data set
1	Considering altitude	4628	1157
2	Disregarding altitude	4628	1157

The next section explains how classifying the handover data into a degree of congestion is done.

3.6.4 Data Classification

There are steps to determine the degree of road traffic congestion from the collected HO data. In the study, ten parameters were considered to create data for ANN modeling. For preprocessing, the dataset was first randomized and then divided into two data sets, the first

data was taken as a training set, and the second one used for test purpose. The percentage split of the data set is 80% of the data set for training, and the reaming 20% for testing. Classification is done based on LOS primary, which is based on the speed of the test vehicle, which can be found from the DT test log file as well as AADT for the significant intersection, Roads intersection with less than 15000 AADT is considered to be moderately congested, whereas greater than 15000 AADT is deemed to be overcrowded, and roads with the level of services with consideration of speed value greater than 13 mph moderate congested and less than this value congested road[18].

In this chapter we have seen, in general, the materials and methods used for the research, the machine learning types, the methods applied for the experiment namely the MLP ANN, how the data collection process was done including parameter selection and data preprocessing.

Chapter 4 Modeling and Experimentation

In Chapter 3 we have seen the materials and methods used for the research, and in this chapter, the experiment method, the performance measures considered for the methods applied as well as model building and result in a discussion are presented.

4.1 Experiment Method

The ANN algorithm is chosen based on the nature of the problem, the complexity of the problem can be managed in neural network and recommended and guided by related works researchers as compared with other similar methods used before. The technique of classification in the neural network used is a supervised one. For training, the selected algorithm percentage split and cross-validation training method has been used. Figure 4-1 shows the process of ANN development which is the data collection, data pre-processing, dividing the data set, develop model, training the model and test it.

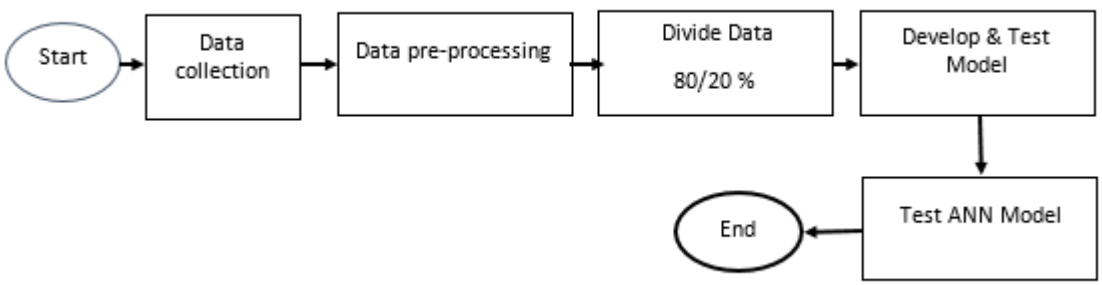

Figure 4-1. Process for development of ANN for estimating road traffic congestion[30]

4.2 Model Building

This section describes training the ANN-MLP, and the MLP has a promising performance for road traffic estimation problems. The developed model was trained using two data sets, which are by considering the altitude feature and ignoring this feature as well as on a moderate elevation difference. For the weekdays, the data set used is from drive test data.

The architecture of the neural network, which is the MLP, consisted of an input layer, one hidden layer, and the output layer, and all are entirely interconnected. The input layer fed

four/three variables, to the hidden layer, which comprised 4/3 nodes. The hidden nodes and the output node all utilized a tan-sigmoid activation function. The target values ranged from 2-3 (2 for moderately congested road traffic level indicator and 3- congested road traffic level indicator). Figure 4-2 shows the network configuration of the experiment.

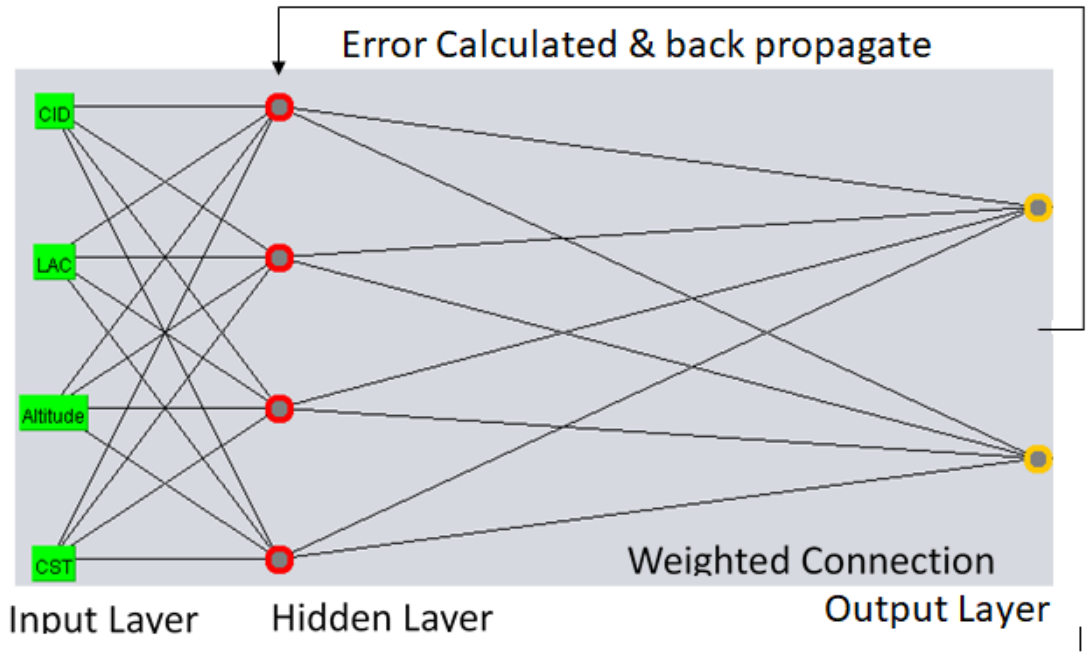

Figure 4-2. Network Configuration

ANN has the capability to distinguish the non-linear association between input and output features and can give generalized solutions to estimate traffic congestion. Multi-Layer Perceptron (MLP) is one of the famous network structure of ANN with an extra layer called a hidden layer. MLP can be used to resolve diverse problems because of the non-linear feature of activation function between its layers of processing elements. The selection of activation function plays an essential role in the performance of a neural network. The error is calculated at each epoch by comparing the computed output of each input with the expected output. Backpropagation is a widely used technique to propagate the error[16].

The objective of the ANN optimization technique is to decrease the error in the training and testing phase. Independently, each processing element is first assigned a random weight. There is no restriction on the choice of few, or many input variables in ANN modeling. The choice of many output and input variables depends on the type of problem issue dealing with. Furthermore, the weight's parameters are modified by propagating the prediction error backward. Parameters like the number of input variables, number of hidden layers, activation function, and learning rate play an important role in the neural network.

In this study, a multilayer perceptron network has been used for the estimation of traffic congestion for a short period. For the development of the ANN model, 6090 data samples have been taken, each of which contained 3/4 features, i.e., CID, CST, Altitude, and LAC. The development and implementation of the ANN model were done with Weka 3.9. The best performing neural network structure is obtained by getting the best values of network parameters for the training and the testing. Seven different ANN models have been developed to train on the dataset prepared. The description of the seven developed models has been presented in Table 4-1. From this table, it is different ANN networks architectures for road traffic estimation, and it indicates that neural network with 4 hidden neurons gives the best prediction result and mean absolute error (MAE) were used to evaluate the performance of predicted results comparing with the rest of the configuration in general.

Table 4-1. The Performance of Developed Model

ID	Learning rate	Momentum	Accuracy
1	0.001	0.2	87.9
2	0.01	0.2	92.6
3	0.09	0.2	92.2
4	0.3	0.001	92.5
5	0.3	0.4	92.4
6	0.3	0.9	92.6
7	0.3	0.99	73.2

The performance of the developed model shown in Table 4-2, Table 4-4, and Table 4-3, which compares the accuracy result found when the altitude factor is considered and not considered.

Table 4-2. Model Performance in Weekdays Considering Altitude

Training model	Result	
	F-Measure	Accuracy
Percentage split	93.5	93.5

Table 4-3. Model Performance in Weekdays Disregarding Altitude

Training model	Result	
	F-Measure	Accuracy
Percentage split	81.5	81.5

Table 4-4. Model Performance with moderate elevation difference (considering altitude)

Training model	Result	
	F-Measure	Accuracy
Percentage split	93	93

4.3 Performance Measure

The model evaluation performed with different evaluation metrics, the suggested metrics for the experiment, comprises Precision, recall, F-measure, ROC or PR curve, Accuracy.

I. Precision

Precision shows the rate with which a fraction of those predicted positive is positive. It usually is not so descriptive of performance independently. For the reason that it only emphasizes the positive instances, it notifies nothing about the negative cases. Subsequently, there is usually a significant trade-off between precision and recall; it is more significant when used in comparison to recall.

Precision described mathematically as in Equation 4-1.

$$\text{Precision} = \frac{TP}{\text{TP} + \text{FP}} \qquad \text{(Equation 4-1)}$$

When TP is true positive count data and FP for false positive count data.The precision result found for the experiment is shown in Table 4-5 and the training result is of weighted average.

Table 4-5. Precision result for the experiment

ID	Model	Precision
1	Considering altitude	92.5
2	Disregarding altitude	82.7

II. Recall

Recall, also known as True Positive Rate, measures the proportion of actual positives which are appropriately identified as positive by the classifier. Frequently it is considered together with precision; meanwhile, there exists a visible tradeoff between them. Recall expressed mathematically as in Equation 4-2

$$\text{Recall} = \frac{TP}{\text{TP} + \text{FN}} \qquad \text{(Equation 4-2)}$$

Where FN is the false negative data and TP the true positive data.

The recall result found for the experiment is shown in

Table 4-6 and the training result is of the weighted average.

Table 4-6 Sample. Recall result for the experiment

ID	Model	recall
1	Considering altitude	92.6
2	Disregarding altitude	82.1

III. F-Measure

It is a measure that combines recall besides precision. It is the mean of precision and recall. It can be expressed mathematically as in Equation 4-3.

$$F - Measure = 2.\frac{\text{Precision. Recall}}{\text{Presision} + \text{Recall}} \quad \text{(Equation 4-3)}$$

The F-measure result found for the experiment is shown in Table 4-7 and the training result is of weighted average.

Table 4-7. F-Measure result

ID	Model	F-measure
1	Considering altitude	92.5
2	Disregarding altitude	81.9

IV. The Receiver Operating Characteristics (ROC)

The receiver operating characteristics curve displays the performance of a model used for the experiment, it is a graph of true positive rate versus false-positive rate. Proficiency of the classifier measured by the area that categorizes the input data properly. The balance between true positive and false negative can be seen on the ROC curve. Figure 4-3 shows the ROC curve of the experiment classifier with an area of under the curve is 0.94.

Any classifier that appears in the lower right triangle performs inferior to the classifier that appears in the upper left triangle[31].

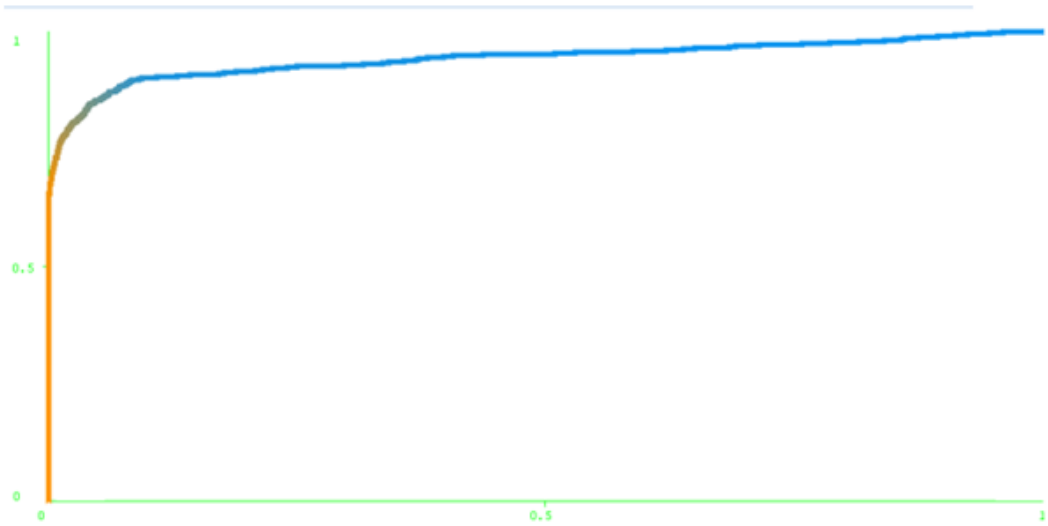

Figure 4-3. ROC of the Classifier

V. Accuracy

Accuracy is one of the most commonly used measures for the classification performance, and it is defined as a ratio between the correctly classified samples to the total number of samples as follows in Equation 4-4[31].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{FP} + \text{FN} + \text{TP} + \text{TN}} \qquad \text{(Equation 4-4)}$$

When TP is the true posive , TN true negative, FP false negative, FN false negative data. The accuracy result found in training is shown in Table 4-8.

Table 4-8. Accuracy Result

ID	Model	Accuracy
1	Considering altitude	92.588
2	Disregarding altitude	82.11

As we observed on the performance measures considered , the methods selected shows promising result.

Chapter 5 Result and Discussion

In Chapter 4, we discussed the experiment method as a summary, the performance measures considered for the model evaluation and model building. This chapter presents the result and discussion part.

The main concern of this study is to observe the impact of altitude in the process of road traffic estimation model which is built with ANN/MLP and Table 5-1 and Figure 5-1 shows the accuracy found with three test case considered, which are by regarding and disregarding the altitude feature for the weekdays and the third is when there is moderate elevation difference on the data set as we can see from the result that the accuracy is better when with inclusiveness of elevation, and when the difference of the altitude in the data set increases the accuracy also have slight improvement up to 1% accuracy improvement, as observed for the field test when there is elevation variety the occurrence of handover observed frequently comparing with flat elevation, technically one of the possible reason can be line of sight of the serving cell and dedicated strong signal.

Table 5-1. Comparison of Models Using Different Data Set

ID	Model	Accuracy
1	Weekday considering altitude	92.5
2	Weekday Disregarding altitude	82.1
3	Addis Ababa(2019) + with moderate elevation difference	93.5/94

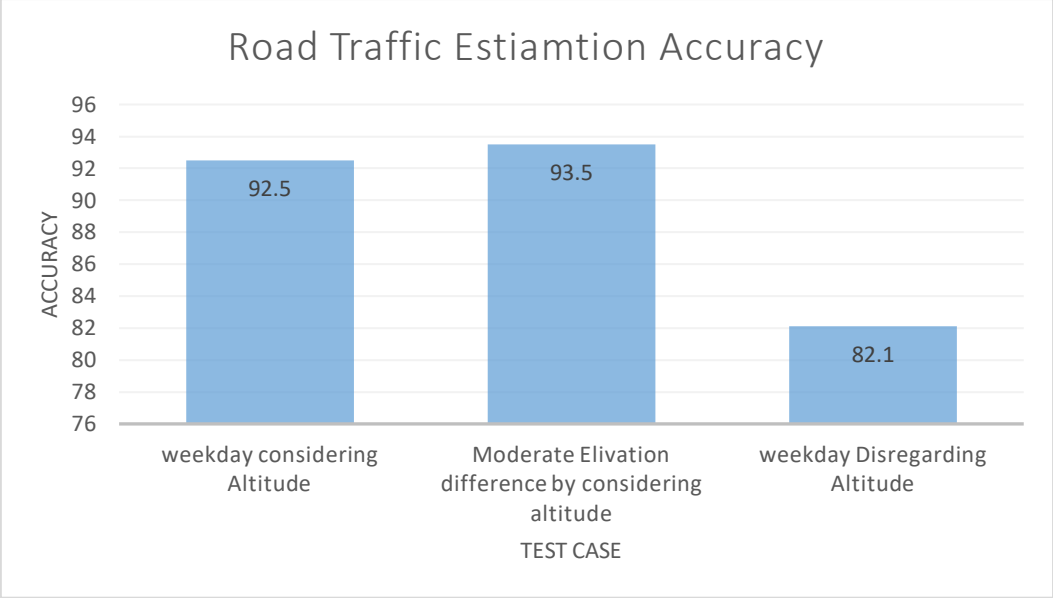

Figure 5-1. Road traffic estimation Accuracy with a different case of elevation

With the new feature, the accuracy shows improvement which increases the true positive count (TP).

Disregarding and considering the altitude feature with a respective confusion matrix of the test set shown in Table 5-2 and Table 5-3 as we can see from the table the true positive count rise in case of considering the new feature, that makes the accuracy to show improvement.

Table 5-2.Confusion matrix disregarding the altitude feature

A	B	Classified as
223	118	A=2
108	708	B=3

Table 5-3.Confusion matrix regarding the altitude feature

A	B	Classified as
228	59	A=2
34	782	B=3

The main research question of the research is that in what level the new parameter altitude will impact the estimation of road traffic using ANN considered in the experiment, therefore comparing the result found with the primary objective, elevation has a visible impact on the estimation model performance used in the experiment and shows 10% accuracy improvement when considering the altitude feature. The technical reason for the effect can be:

By considering the effects of altitude , the estimation difference is insignificance when the test area elevation is small and improved when there is variation of the altitude. Technicaly when there is a site that serves long possible coverage disregarding other factors, even if there is no traffic congestion on the road there will not be HO, this shows long serving time of the cell and the estimation lead to congested, and the contrary way there will be a site that has minimum coverage area related to topology so that the cell serving time will be less that lead to the estimation to free.

Most of the related researchers perform their study using four main parameters, which are CID, Time, CDT, and LAC. Which means without considering altitude as a factor, in this research we perform the experiment by considering and ignoring altitude impact on the estimation of road traffic information. we have found that it has significant effect on the estimation model especially when there is big variation on the elevation of test road considered.

In the study, we planned three data set classification on weekdays considering and disregarding the altitude parameter, and also considering moderate elevation difference, and based on the nature of the problem type, we selected the ANN/MLP method for model building and testing of the experiment.

The size of cell coverage areas varies from one location to another. To address this issue, related works are done by including parameters of location area code (LAC) and Cell Identification (CID) and also on this experiment that provided by the base stations to uniquely identify the cells and associate them with geographic areas. The neural network can then learn the comparative size of a given cell from the relationship between CST and user judgments of congestion levels[13].

Table 5-4 and Figure 5-2 show the result found comparing with other related research works.

Table 5-4. Estimation accuracy with related work

ID	Previous work	Accuracy (%)
1	Tiland(2007)	79
2	Portugal(2013)	82
3	UK(2017)	87
4	Addis Ababa(2019) +	92.5
5	Addis Ababa(2019) -	82.1
6	Addis Ababa(2019) + with moderate elevation difference	93.5/94

Note:

+ Considering the altitude parameter
- Disregarding the altitude parameter

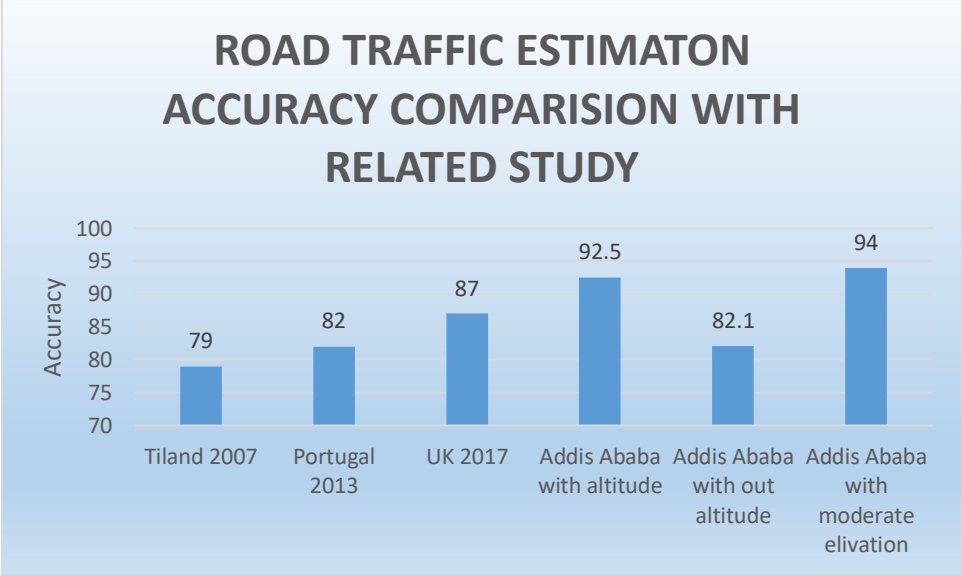

Figure 5-2. Road traffic estimation Accuracy comparison with related work

The result found without considering the altitude parameter is almost similar to other research works done. Due to different factor it is not visible to comare directly , whereas comparing with the new parameter, there is a significant variation of the level of accuracy.

The improvement in the estimation accuracy can be on, the geographical location of the experiment area as well as the data set of the technology considered some of the researchers mentioned that the data set used 2G for their experiment.

In general, road traffic estimation result found using handover data of 3G and MLP method by considering the altitude parameter as a factor shows an improvement especially when there is visible elevation difference for the experiment area.

Chapter 6 Conclusion and Future work

this chapter includes conclusion and recommendation as future work.

6.1 Conclusion

As discussed in chapters three and four, the study uses the backpropagation ANN-MLP approach model for road traffic estimation with mixed traffic and the arterial road of Addis Ababa. The significant advantage of the back-propagation neural network is that it calculates the prediction error and propagates back to the previous layers to modify the weights, resulting in better prediction accuracy with more training. The training is stopped after a certain number of epochs and when there are no further improvements in estimation accuracy.

As the experiment, the result found to confirm that the parameter altitude has a significant impact on the process of road traffic estimation since node B coverage is affected by elevation. Based on the result found, we can conclude that considering the new feature for evaluation will increase the level of estimation accuracy by 10% for information users, which can lead to other possible options for road selection and can minimize the traffic congestion level of the city to certain level.

The study is certainly capable of providing the right solution for short term traffic prediction for arterial roads with mixed traffic conditions in Addis Ababa. However, the dataset used for this study is on a limited portion of the arterial road. To enhance the study, using more informative data about traffic congestion conditions including seasonality and for roads that are free from congestion is essential. The data collection procedure requires to collect data on peak hour or busy hour, the time that users use to work and from work to home.

Road traffic estimation has potential applications in traffic management and for longer-term performance monitoring of the road network. Compared with available alternatives, mobile phones offer attractive characteristics as traffic inquiries.

Implementation of road traffic estimation can enable, transport system performance and for travelers to facilitate informed decisions by travelers about the route.

6.2 Future work

As a recommendation for future work, the classification of road congestion levels considered in this research are moderately congested and congested on with two classification, the study does not include a free congestion level due to data sample unavailability, so it will be better to see with additional test data. This research uses HOD as an input for the road traffic estimation, and there is also the possibility to study using call detail record data (CDR), the future scope of this study would be to work on mentioned limitations of the research and provide a better solution.

Reference

[1] K. Adam and A. Vakali, “Real-time Traffic Prediction using Multiple Data Sources,” 2018.

[2] T. Wondwossen, “Assessing & Quantifying the Level of Traffic Congestion at Major Intersections in Addis Ababa,” Addis Ababa University, 2011.

[3] G. Dereje, “UMTS Traffic Model Using ANN: The case of Addis Ababa, Ethiopia,” 2017.

[4] C. P. & T. Truong, “3G (UMTS) HANDOVER ISSUES.” [Online]. Available: https://www.scribd.com/presentation/351608976/3g-Handover-Issues. [Accessed: 11-Nov-2019].

[5] S. Bekele, “Cell Outage Detection Through Density-based Local Outlier Data Mining Approach :In case of Ethio telecom UMTS Network,” no. November, 2018.

[6] H. J. Von Schmoeger, *umts-the-fundamentals*. 2001.

[7] M. Wolde, “An Overview of Addis Ababa Transport System,” Addis Ababa University, 2016.

[8] M. Murad, “Costing Road Traffic Accidents In Ethiopia,” Addis Ababa University, 2011.

[9] M. G. Demissie, “Traffic Volume Estimation Through Cellular , Coimbra University,” no. July, pp. 1–16, 2013.

[10] T. Hansapalangkul, P. Keeratiwintakorn, and W. Pattara-Atikom, “Detection and estimation of road congestion using cellular phones,” *ITST 2007 - 7th Int. Conf. Intell. Transp. Syst. Telecommun. Proc.*, vol. 2, no. 1, pp. 143–146, 2007.

[11] M. A. Gebresilassie, “Spatio-temporal Traffic Flow Prediction,” KTH - Royal Institute of Technology, 2017.

[12] W. Hongsakham, W. Pattara-Atikom, and R. Peachavanish, “Estimating road traffic congestion from cellular handoff information using cell-based neural networks and K-means clustering,” *5th Int. Conf. Electr. Eng. Comput. Telecommun. Inf. Technol. ECTI-CON 2008*, vol. 1, pp. 13–16, 2008.

[13] W. Pattara-atikom and R. Peachavanish, “Estimating Road Traffic Congestion from Cell Dwell Time using Neural Network,” *2007 7th Int. Conf. ITS Telecommun.*, pp. 9–11, 2007.

[14] M. R. Singha, “Using Mobile Phone Network for Urban Traffic Management,” vol. 65, no. 2, pp. 12–17, 2013.

[15] F. M. S. C.P.IJ. van Hinsbergen, J.W.C. van Lint, “Short Term Traffic Prediction Models,” no. February, 2014.

[16] B. Sharma, S. Kumar, P. Tiwari, P. Yadav, and M. I. Nezhurina, “ANN based short - term traffic flow forecasting in undivided two lane highway,” *J. Big Data*, 2018.

[17] H. Yared, “Impact of Vehicle Traffic Congestion in Addis Ababa,” Ethiopian Civil Service College, 2010.

[18] R. G. Dowling and B. K. Ostrom, “Highway Capacity Manual 2010,” no. March 2011, 2016.

[19] G. Leduc, “Road Traffic Data : Collection Methods and Applications,” 2008.

[20] A. Galloni, B. Horváth, and T. Horváth, “Real-time Monitoring of Hungarian Highway Traffic from Cell Phone Network Data,” vol. 2203, pp. 108–115, 2018.

[21] P. Bertagna, “How does a GPS tracking system work,” 2010. [Online]. Available: https://www.eetimes.com/document.asp?doc_id=1278363&page_number=2. [Accessed: 21-Nov-2019].

[22] M. Bachani, “Intelligent Transportation System International Journal of Scientific & Engineering Research, Volume 4, Issue 1,” 2013.

[23] Thomas H. Davenport, “Machine Learning,” 2019. [Online]. Available: https://www.sas.com/ru_ru/insights/analytics/machine-learning.html. [Accessed: 11-Nov-2019].

[24] P. Gaur, “Neural Networks in Data Mining,” pp. 1449–1453, 1956.

[25] R.Rojas, “Perceptron Learning -Neural Networks,” Berlin: Springer-Verlag, 1996.

[26] K. Verma, “An Insight to Soft Computing based Defect Prediction Techniques in Software,” no. October, 2016.

[27] C. Zhang and Q. Yang, “Data Preparation for Data Mining.,” no. May 2014, 2003.

[28] K. Hagos, “SIM-Box Fraud Detection Using Data Mining Techniques : The Case of ethio telecom,” Addis Ababa, 2018.

[29] M. A. Hall, “Correlation-based Feature Selection for Machine Learning,” no. April, 1999.

[30] B. T. A. Borja García de Soto, Andreas Bumbacher, Markus Deublein, “Predicting Road Traffic Accidents using Artificial Neural Network Models,” pp. 1–49, 2018.

[31] A. Tharwat, “Applied Computing and Informatics Classification assessment methods,” *Appl. Comput. Informatics*, 2018.

APPENDIX

- Sample Data Set

  - Considering the Altitude Feature

```
@relation 'merged_data_with_new_one-weka.filters.unsupervised.attribute.NumericToNominal-Rlast-weka.filters.unsupervised.attribute.
Reorder-R1,2,3,4,5-weka.filters.unsupervised.instance.Resample-S1-Z80.0-no-replacement-V'
@attribute CID numeric
@attribute LAC numeric
@attribute Altitude numeric
@attribute CST numeric
@attribute HUO {2,3}
@data
52283,1105,2549,160.4,3
35985,1101,2317,292,3
30179,1105,2294,50,2
30178,1105,2244,151,2
30857,1105,2221,210,3
30321,1102,2501,462.38,3
55986,1101,2395,253,3
35984,1101,2316,185,3
31651,1105,2384,193,3
52833,1104,2295,160,2
30126,1104,2265,179,2
32595,1104,2259,138,2
35987,1101,2316,218,3
33762,1105,2379,261,3
36947,1105,2506,20,3
56584,1105,2506,47.2,3
32678,1101,2380,140,3
30123,1104,2301,160,2
30855,1105,2315,283,3
35343,1104,2530,96.4,3
56291,1105,2360,248,3
32282,1105,2369,220,2
50881,1105,2315,124,2
30885,1105,2315,196,2
50175,1105,2315,151,2
33678,1104,2301,201,3
32594,1104,2296,129,2
33762,1105,2396,157,3
57286,1105,2520,32.8,3
```

- Disregarding the Altitude Feature

```
@relation 'merged_data_with_new_one-weka.filters.unsupervised.attribute.NumericToNominal-Rlast-weka.filters.unsupervised.attribute.
Remove-R3-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4-weka.filters.unsupervised.instance.Resample-S1-Z80.0-no-replacement-V'
@attribute CID numeric
@attribute LAC numeric
@attribute HOFC numeric
@attribute HUO {2,3}
@data
52283,1105,160.4,3
35985,1101,292,3
30179,1105,50,2
30178,1105,151,2
30857,1105,210,3
30321,1102,462.38,3
55986,1101,253,3
35984,1101,185,3
31651,1105,193,3
52833,1104,160,2
30126,1104,179,2
32595,1104,138,2
35987,1101,218,3
33762,1105,261,3
36947,1105,20,3
56584,1105,47.2,3
32678,1101,140,3
30123,1104,160,2
30855,1105,283,3
35343,1104,96.4,3
56291,1105,248,3
32282,1105,220,2
50881,1105,124,2
30885,1105,196,2
50175,1105,151,2
33678,1104,201,3
32594,1104,129,2
33762,1105,157,3
57286,1105,32.8,3
50983,1105,121.2,3
```

- Training Run Information with respect to the data set

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        4285               92.5886 %
Incorrectly Classified Instances       343                7.4114 %
Kappa statistic                          0.8082
Mean absolute error                      0.0979
Root mean squared error                  0.2372
Relative absolute error                 24.5732 %
Root relative squared error             53.149  %
Total Number of Instances             4628

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.817    0.033    0.903      0.817    0.858      0.810    0.942     0.928     2
                 0.967    0.183    0.933      0.967    0.950      0.810    0.942     0.959     3
Weighted Avg.    0.926    0.142    0.925      0.926    0.925      0.810    0.942     0.951

=== Confusion Matrix ===

    a    b   <-- classified as
 1038  232 |    a = 2
  111 3247 |    b = 3
```

```
=== Summary ===

Correctly Classified Instances        3800               82.1089 %
Incorrectly Classified Instances       828               17.8911 %
Kappa statistic                          0.5401
Mean absolute error                      0.2623
Root mean squared error                  0.3645
Relative absolute error                 65.8524 %
Root relative squared error             81.6755 %
Total Number of Instances             4628

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.637    0.109    0.688      0.637    0.661      0.541    0.846     0.626     2
                 0.891    0.363    0.866      0.891    0.878      0.541    0.846     0.933     3
Weighted Avg.    0.821    0.293    0.817      0.821    0.819      0.541    0.846     0.849

=== Confusion Matrix ===

   a    b   <-- classified as
 809  461 |    a = 2
 367 2991 |    b = 3
```