

ADDIS ABABA UNIVERSITY
COLLEGE OF TECHNOLOGY AND BUILT ENVIRONMENT
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

Machine Learning-based Congestion Prediction in Long Term Evolution Networks

By:
Atsedemariam Gebeyehu

Advisor:
Dr. -Ing. Dereje Hailemariam

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Electrical and Computer Engineering.

May 2026
Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY

COLLEGE OF TECHNOLOGY AND BUILT ENVIRONMENT

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

COMMUNICATION ENGINEERING PROGRAM

Machine Learning-Based Congestion Prediction in Long Term Evolution Networks

By:

Atsedemariam Gebeyehu Agegnehu

Approval by the Board of Examiners

Chairman, School Graduate Committee	Signature	Date
_____	_____	_____

Committee

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

Abstract

In Ethiopia, LTE networks are currently the most widely used technology for mobile broadband services. This is mainly because LTE infrastructure has already been deployed in many areas and supports a large number of mobile users. Although 5G services have started in some areas of Addis Ababa, most mobile data users still depend on LTE networks. The continuous growth of mobile data consumption has increased the load on LTE cells, particularly in locations where many users access the network at the same time. When traffic becomes high, LTE cells may experience lower throughput and higher delay, which can negatively influence the QoS experienced by users. These problems are commonly observed in highly populated areas such as Addis Ababa, where network traffic changes depending on user activities throughout the day.

This paper develops a Random Forest-based model for predicting LTE throughput congestion using network performance measurements.. The model uses different LTE Key Performance Indicators (KPIs) related to radio signal quality, resource usage, and traffic conditions. In addition, feature importance analysis is applied to determine which KPIs have a stronger influence on throughput degradation and congestion.

Random Forest regression was chosen because LTE performance depends on several KPIs at the same time. The algorithm can learn the relationship between these KPIs and throughput variation. The results show that the model can estimate LTE throughput conditions and identify important network parameters related to congestion. The performance of the developed model is measured using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). The results show that the model can estimate LTE throughput conditions and identify important network parameters related to congestion.

This research is based on historical LTE KPI data collected from selected eNodeB cells in the Bole Bulbula area of Addis Ababa. Future research may improve the model by including real-time KPI measurements and data collected from additional LTE sites or other wireless networks."

Keywords: LTE, Throughput Prediction, Congestion Analysis, Random Forest Regression, Machine Learning, eNodeB, KPI Analysis, Feature Importance

Declaration

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

Atsedemariam Gebeyehu Agegnehu _____

Name Signature

Addis Ababa _____ / _____

Place Date of Submission

The above candidate has carried out research for the master's thesis under my supervision.

Dr. -Ing. Dereje Hailemariam _____

Name Signature

Acknowledgment

First and foremost, I would like to thank the Creator of heaven and earth for the strength and health provided to me. I am deeply grateful to His Holy Mother St. Mary for her continuous guidance and protection throughout this entire process.

I would also like to express my deepest gratitude to my advisor, Dr. -Ing. Dereje Hailemariam, for his immense guidance and expertise during the making of this study. Without his consistent motivation, professional support, and academic guidance, this research work would not have been possible.

A special thanks goes to my boss, Mr. Tesfamichael Gebretsadik, whose humble personality and genuine support provided me with the necessary environment to finish this thesis. His encouragement was truly valuable throughout this journey.I would also like to sincerely thank my colleague, Mr. Wondmagegn Wana, for his valuable support, insightful discussions, and encouragement throughout the development of this thesis.

Finally, I would like to thank my family and all of my friends for their unwavering support and encouragement. This achievement belongs to all of us we did it, and we did it good.

List of Acronyms

3GPP	3rd Generation Partnership Project.
AMC	Adaptive Modulation and Coding.
API	Application Programming Interface.
BER	Bit Error Rate.
BLER	Block Error Rate.
BPSK	Binary Phase Shift Keying.
CCE	Control Channel Element.
CNN	Convolutional Neural Network.
CQI	Channel Quality Indicator.
DL	Downlink.
E-RAB	Evolved Radio Access Bearer.
eNB	Evolved Node B.
EUTRAN	Evolved Universal Terrestrial Radio Access Network.
GBM	Gradient Boosting Machine.
GBR	Gradient Boosting Regression.
GPS	Global Positioning System.
IQR	Interquartile Range.
KNN	K-Nearest Neighbors.
KPI	Key Performance Indicator.
LSTM	Long Short-Term Memory.
LTE	Long Term Evolution.

MAE	Mean Absolute Error.
MCS	Modulation and Coding Scheme.
MIMO	Multiple-Input Multiple-Output.
ML	Machine Learning.
NFV	Network Function Virtualization.
NMS	Network Management System.
ODL	OpenDaylight.
OFDM	Orthogonal Frequency Division Multiplexing.
PGW	Packet Data Network Gateway.
PRB	Physical Resource Block.
PRS	Performance Reporting System.
QAM	Quadrature Amplitude Modulation.
QoE	Quality of Experience.
QoS	Quality of Service.
QPSK	Quadrature Phase Shift Keying.
RAN	Radio Access Network.
RF	Random Forest.
RL	Reinforcement Learning.
RMSE	Root Mean Square Error.
RNN	Recurrent Neural Network.
RRC	Radio Resource Control.
RSRP	Reference Signal Received Power.
RSRQ	Reference Signal Received Quality.
RSSI	Received Signal Strength Indicator.
RTT	Round Trip Time.
SDN	Software Defined Networking.
SGW	Serving Gateway.
SINR	Signal-to-Interference-plus-Noise Ratio.

SNR	Signal-to-Noise Ratio.
SVM	Support Vector Machine.
SVR	Support Vector Regression.
TA	Timing Advance.
TCP	Transmission Control Protocol.
UE	User Equipment.
UL	Uplink.
XGBoost	Extreme Gradient Boosting.

Contents

Abstract i

Declaration ii

Acknowledgment iii

List of Tables x

List of Figures xi

1 Introduction 1

1.1 Background . . . 1
1.2 Statement of Problem . . . 4
1.3 Literature Review . . . 5
1.4 Objectives . . . 9
1.4.1 General Objective . . . 9
1.4.2 Specific Objectives . . . 9
1.5 Methodology . . . 9
1.6 Scope and Limitation . . . 11
1.6.1 Scope of the Thesis . . . 11
1.6.2 Limitation of the Thesis . . . 12
1.7 Outline of the Thesis . . . 12

2 LTE Throughput 14

2.1 LTE Network Overview . . . 14
2.1.1 LTE Radio Access Network . . . 15
2.1.2 Evolved Packet Core . . . 15
2.1.3 Packet Data Network . . . 16
2.2 Physical Layer Aspects of LTE . . . 17
2.2.1 Scalable Bandwidth and Single-Carrier Limits . . . 17
2.2.2 Downlink and Uplink Multiple Access Schemes . . . 17
2.2.3 Time-Frequency Resource Grid . . . 18
2.2.4 Baseline MIMO Antenna Configurations . . . 19
2.2.5 Adaptive Modulation and Coding . . . 19
2.2.6 Physical Channel Mapping and Signaling . . . 20
2.3 LTE Network Throughput Measurements . . . 21

2.3.1 Direct Measurements . . . 21
2.3.2 Indirect Measurements and Inference . . . 21
2.3.3 Challenges in Throughput Measurement . . . 22
2.4 Quality of Service Impacts of Network Load and Congestion . . . 22
2.4.1 Degradation of User-Plane Throughput . . . 23
2.4.2 Buffer Overflow and Packet Loss . . . 23
2.4.3 Delay Variation and Packet Arrival Inconsistency . . . 24
2.4.4 Application-Specific Performance Vulnerabilities . . . 24
2.4.5 Systemic Decline in User Satisfaction . . . 24

3 Machine Learning Models for LTE Throughput Prediction 25
3.1 Machine Learning in Wireless Networks . . . 25
3.2 Random Forest Regression . . . 27
3.2.1 Decision Tree Ensemble Learning . . . 27
3.2.2 Bootstrap Aggregating . . . 29
3.2.3 Feature Importance Analysis . . . 29
3.3 XGBoost Regression . . . 29
3.3.1 Regularization and Optimization . . . 31
3.4 Deep Learning Models for Throughput Prediction . . . 31
3.4.1 Long Short-Term Memory (LSTM) . . . 32
3.4.2 Hybrid CNN-LSTM Architecture . . . 33
3.5 Reinforcement Learning Approaches . . . 34
3.6 Justification for Selecting Random Forest . . . 34

4 Proposed Random Forest Model and Data Preparation 36
4.0.1 Proposed Model Architecture and Hyperparameters . . . 36
4.1 Data Collection and Preparation . . . 38
4.1.1 Site Selection . . . 38
4.1.2 Data Sources . . . 39
4.1.3 Data Cleaning . . . 40
4.1.4 Data Normalization . . . 40
4.2 Cyclical Feature Encoding for Temporal Attributes . . . 42
4.3 Correlation Analysis . . . 43
4.4 Model Configuration . . . 47

5 Results and Discussion **48**
5.1 Throughput Forecasting Analysis 48
5.2 Feature Importance Analysis 49
5.3 Residual and Error Distribution 51
5.4 Target Variable Correlation Heatmap 52
5.5 Binary Congestion Classification Performance 53
5.6 Model Evaluation Metrics and Performance Analysis 55

6 Conclusion and Recommendations **56**
6.1 Conclusion 56
6.2 Recommendations 56
6.3 Future Work 57

List of Tables

1.1 Summary of Related Works on LTE Throughput Prediction and Congestion Analysis 8
4.1 Eliminated Missing Values from Each Month's Data set 40
4.2 Summary of Proposed Random Forest Model Parameters 47
5.1 Feature Importance Analysis for Congestion Prediction 51
5.2 Performance Evaluation for Congestion Detection Model 54
5.3 Model Accuracy 55

List of Figures

1.1 Monthly PS traffic trends for 3G, 4G, and 5G networks from May 2023 to April 2026 . . . 1
1.2 Distribution of Congested LTE Cells Across Zones in Addis Ababa . . . 3
1.3 Methodology . . . 11
2.1 LTE Network Architecture . . . 15
2.2 Orthogonal Frequency Division Multiplexing . . . 18
2.3 Physical Resource Block . . . 19
2.4 SNR Versus Throughput for Different Modulation Schemes . . . 20
2.5 Throughput Congestion . . . 23
3.1 Decision Tree Ensemble in Random Forest Model . . . 28
3.2 XGBoost vs. Random Forest . . . 30
3.3 Long Short-Term Memory Neural Networks . . . 32
4.1 Sites with Higher Throughput at Bolebulbulla . . . 39
4.2 Average Throughput of The 17 Sites for July/2024 Data . . . 39
4.3 Kendall's Tau Correlation Output . . . 43
5.1 Model Throughput Forecast Test . . . 48
5.2 Actual vs. Predicted Throughput Correlation . . . 49
5.3 Top 10 Feature Importance for Throughput Prediction . . . 50
5.4 Prediction Error Distribution . . . 52
5.5 KPI Linear Correlation Heatmap with Target Throughput . . . 53
5.6 Binary Congestion State Prediction Output . . . 54

1 Introduction

This chapter introduces the main objective of this thesis, which is to develop a machine learning-based throughput congestion prediction model for Long-Term Evolution (LTE) networks. This chapter also provides a detailed review of related literature's and an overview of the thesis structure. It further identifies existing research gaps in today's congestion prediction methodologies. Finally, the chapter outline the scope and limitations of the study.

1.1 Background

Fourth Generation (4G) Long-Term Evolution (LTE) networks remain the most widely used mobile broadband technology in Addis Ababa due to their existing coverage, large subscriber base, and compatibility with commonly used mobile devices. Although 5G services have been introduced in selected areas of the city, their deployment is still limited compared with LTE. The availability of 5G services depends on network coverage and user access to 5G-enabled devices. Therefore, LTE continues to support the majority of mobile broadband users and remains an important technology for providing mobile data services in Addis Ababa.

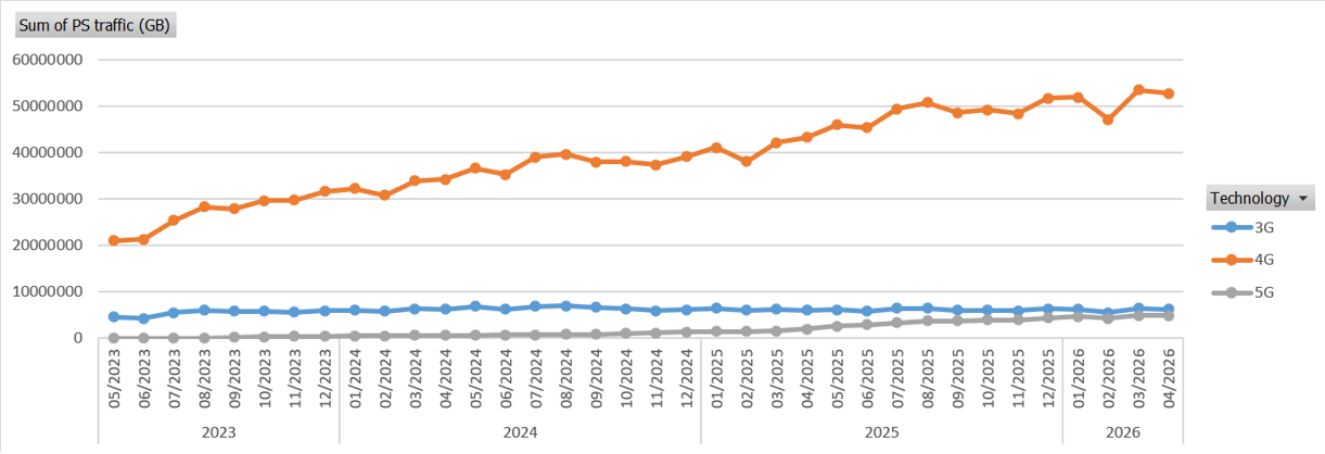


Figure 1.1: Monthly PS traffic trends for 3G, 4G, and 5G networks from May 2023 to April 2026[1]

The continuous growth in smartphone usage and mobile data consumption has increased traffic demand on LTE networks, Results to frequent throughput congestion and reduced Quality of Service in densely populated areas. Monthly Packet-Switched (PS) traffic volume across 3G, 4G, and 5G technologies (Figure 1.1) from May 2023 to April 2026, highlighting higher traffic in 4G and a gradual increase in 5G usage over the period.

As network traffic increases, LTE cells may become congested when available radio resources are no longer sufficient to serve the growing number of active users and data sessions. Network congestion causes several performance issues like slows down transmission speeds, increases delays, and triggers repeated attempts to send data, all of which negatively affect the quality of service. Several LTE Key Performance Indicators (KPIs) highly influence throughput performance either directly or indirectly. Physical Resource Block (PRB) utilization reflects the level of radio resource consumption within a cell, where excessive utilization may lead to severe congestion and throughput degradation. Signal quality indicators such as Signal-to-Interference-plus-Noise Ratio (SINR), Reference Signal Received Power (RSRP), and Reference Signal Received Quality (RSRQ) affect modulation efficiency and transmission reliability, thereby influencing achievable data rates. In addition, Block Error Rate (BLER), Channel Quality Indicator (CQI), and Control Channel Element (CCE) utilization impact retransmission behavior, signaling efficiency, and resource allocation within the network. Accessibility and retainability KPIs, including E-RAB Setup Success Rate and RRC Congestion Ratio, also indirectly influence throughput performance by affecting connection establishment and session stability. Therefore, continuous monitoring and analysis of these KPIs are essential for understanding LTE congestion behavior and developing machine learning-based throughput prediction model[2].

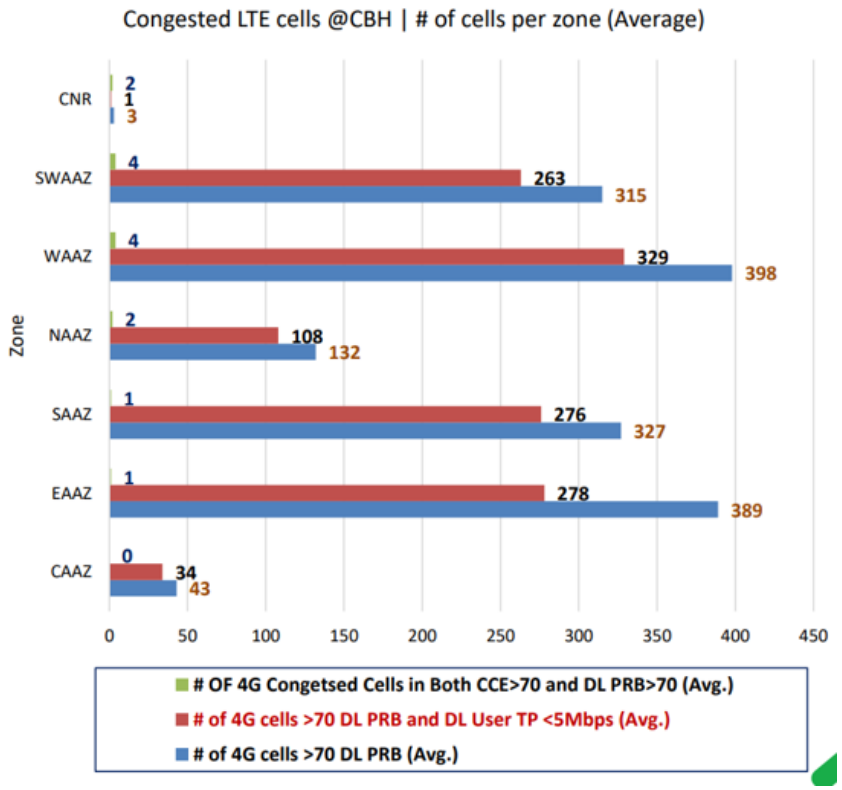


Figure 1.2: Distribution of Congested LTE Cells Across Zones in Addis Ababa[3]

Predicting congestion in ethiotelecoms LTE networks is essential, as it allows the operators to identify performance degradation before it actually impact the user. Traditional network management the one currently performed in ethiotelecom is mostly reactive; it usually alerting engineers to a problem once customers are already experiencing poor service. ML help operators to shift from the proactive approach to actively anticipate congestion before it actually happens. This transition helps to maintain stable throughput, improves overall Quality of Service (QoS), and ensures network resources are utilized as efficiently as possible.

Machine learning has become an effective approach for predicting network throughput [4]. By using historical network data, Ethiotelecom can improve from only monitoring network conditions to predicting possible congestion before it affects users.

ML algorithms such as Random Forest and Gradient Boosting Machines (GBMs) are commonly used because they provide good prediction accuracy and can handle complex network data with less risk of overfitting and other algorithms like deep learning meth-

ods ; Long Short-Term Memory (LSTM) networks are useful for learning time-based changes in network traffic. These approaches help to develop better methods for network performance monitoring and optimization.

This thesis focuses on developing a Random Forest-based prediction model to study the relationship between LTE network parameters and throughput performance. LTE networks support most of the mobile broadband services in Addis Ababa, and the continuous growth of user demand can create congestion problems that reduce network performance. Using real LTE network data provides an opportunity to analyze these challenges under practical operating conditions. Besides throughput prediction, the model also identifies the main KPIs that influence performance changes, which can support network engineers in making better decisions for congestion management.

1.2 Statement of Problem

LTE networks are used to serve the majority of mobile broadband data traffic in densely populated regions such as Addis Ababa where 4G access is widely accessible than Emerging 5G networks. Ethio Telecom's network frequently experiences congestion, particularly in the West Addis Ababa zone, as a result of an increase in customers, the widespread usage of smart devices, and the popularity of data-intensive video streaming. Users are irritated by these LTE cells' slower data speeds, higher latency, and lost packets during demand spikes. The existing systems operate on Network Management Systems (NMS), which tend to alert engineers only when there is a problem; in essence, the user's link has already been affected. Although statistical techniques have been employed to predict future traffic loads, they are prone to having difficulty in understanding the complex relationship between network metrics and real link speeds.

The NMS currently deployed at Ethio Telecom has the capability of tracking real-time metrics such as throughput, utilization of physical resource blocks, signal strength, and network load. It has been challenging to manually manage these metrics by using static alerts, especially as the patterns have become increasingly volatile. Through machine learning, these patterns can be detected in historical data, enabling prediction of future problems.

Most literature written about this topic relies on a very narrow selection of metrics, which include only PRB utilization without considering other parameters like SINR, BLER, CCE, or availability. However, the need for analysis using a wider variety of KPIs to examine network throughput in the particular circumstances of Addis Ababa has not been sufficiently explored before. This thesis attempts to fill this gap by constructing a Random Forest regression algorithm based on the data collected at Bole Bulbula's eNodeB. In addition to predicting future speeds, this model identifies the variables responsible for congestion, allowing the engineer to solve the problem by targeting its sources directly.

1.3 Literature Review

LTE throughput prediction is an important aspect of network performance analysis because throughput variation is directly related to user experience and congestion conditions. Recent studies have explored different prediction approaches using LTE KPIs collected from user measurements and operator Network Management Systems (NMS). Researchers have proposed various approaches for throughput estimation and prediction using both user-side measurements and data collected from operator Network Management Systems (NMS).

L. Alho. [5] nvestigated LTE throughput prediction using measurements collected from public transit buses in Kingston, Ontario. The dataset was collected from repeated LTE measurements during different trips and included radio parameters such as RSRP, RSRQ, RSSI, and SNR. The dataset included LTE parameters such as RSRP, RSRQ, RSSI, and SNR, along with contextual information including GPS coordinates and vehicle speed. Several machine learning algorithms were evaluated using a 70/30 training and testing approach, and the Random Forest model achieved the highest prediction result among the tested models with an R^2 score of 0.78 and the lowest Root Mean Square Error (RMSE). However, the study mainly evaluated prediction accuracy and did not analyze KPI contribution to congestion.

Abolfazli et al. [6] the authors evaluated the accuracy of throughput measurement applications such as Iperf and OOKLA within a 4G WiMAX network environment. The results indicated that both tools provide nearly accurate throughput measurements for

analysis of network bandwidth. The authors suggested that combining these tools could support throughput monitoring analysis.However, these tools focus mainly on user-side measurements and provide limited information about LTE KPIs and congestion causes from the operator perspective.

Eyceyurt et al. [7]This study explored machine learning for predicting LTE uplink throughput by analyzing ten days of data across various international sites. The researchers examined network parameters such as RSRP, RSRQ, and SNR using algorithms like Support Vector Regression (SVR), Random Forest, Decision Trees, K-Nearest Neighbors (KNN), and Gradient Boosting Regression (GBR). The results indicated that the Decision Tree and KNN models were particularly effective, as shown by their R^2 scores. However, the study didn't fully look at how all these different network factors interact with each other to cause slow speeds. This gap highlights a clear need for other models capable of weighing multiple network metrics to better understand what truly causes throughput fluctuations.

Alho et al. [8] proposed a hybrid neural network architecture using supervised and unsupervised learning techniques to identifying low-throughput eNodeB cells. Using real-world LTE monitoring data with limited labeled samples, the objective of the proposed approach is to automate the identification of problematic network cells and reduce manual troubleshooting efforts for network operators. Although deep learning models can achieve high prediction performance, their limited interpretability can make it difficult to understand the contribution of individual KPIs. In LTE environment it is important to understand how and why a model is making a decision, and the complexity of these deep learning techniques so to use it in the live LTE network.

Wang et al. [9] like the previously mentioned study; this paper also mentioned that deep learning-based traffic prediction approaches for mobile communication networks achieve prediction performance improvements compared to traditional statistical techniques. However, several limitations were identified, including high computational requirements, large training data dependency, and the difficulty of interpreting model decisions due to their black-box nature.These limitations make it difficult to identify the specific LTE KPIs responsible for congestion conditions.

Zhang et al. [10] investigated mobile network throughput prediction using ensemble ma-

chine learning techniques and large-scale operational network datasets. The study compared Random Forest, Gradient Boosting, and Artificial Neural Networks for predicting user throughput under varying network conditions. The experimental results showed that ensemble models achieved better prediction performance than individual models. The authors showed that radio quality parameters, such as RSRP and SINR, have an effect on throughput performance. However, their work mainly concentrated on improving prediction accuracy and gave less attention to analyzing KPIs related to congestion.

In a similar study, Nguyen et al.[**?**] developed a data-driven approach for predicting LTE network performance using KPI data collected from the network operator. The study analyzed how throughput changes with different network conditions, including Physical Resource Block (PRB) utilization, the number of active users, and channel quality parameters.The findings showed that resource utilization metrics strongly affect throughput degradation during peak traffic periods. Nevertheless, the proposed approach relied on complex feature engineering processes, which may limit practical deployment in operational networks.

Previous studies show that LTE throughput is influenced by radio quality indicators such as RSRP, RSRQ, SNR, SINR and resource utilization parameters such as PRB utilization and active users. Machine learning methods have been applied successfully for throughput prediction; however, most studies focus mainly on prediction accuracy. Limited research investigates the contribution of multiple LTE KPIs to congestion conditions. Therefore, this thesis develops a Random Forest-based model that predicts LTE throughput and identifies important KPIs affecting congestion using feature importance analysis.

Table 1.1: Summary of Related Works on LTE Throughput Prediction and Congestion Analysis

Reference	**Method / Model**	**KPIs / Data Used**	**Key Findings**	**Limitations**
Alho [5]	Random Forest	RSRP, RSRQ, RSSI, SNR, GPS, Vehicle Speed	RF achieved the best performance with $R^2 = 0.78$ and the lowest RMSE.	Limited analysis of KPI relationships and congestion behavior.
Abolfazli et al. [6]	Iperf, OOKLA	4G WiMAX throughput measurements	Reliable throughput measurement tools for bandwidth analysis.	No LTE KPI analysis or congestion diagnosis capability.
Eyceyurt et al. [7]	SVR, RF, DT, KNN, GBR	RSRP, RSRQ, SNR	DT and KNN achieved strong throughput prediction performance.	Did not investigate interactions among multiple LTE KPIs.
Alho et al. [8]	Hybrid Neural Network	LTE monitoring data	Identified low-throughput cells automatically.	High complexity and limited interpretability.
Wang et al. [9]	Deep Learning Review	Mobile traffic datasets	Deep learning improves prediction accuracy.	Large training data requirements and black-box behavior.
Verma et al. [10]	RF, GB, ANN	RSRP, SINR	Ensemble models achieved high prediction accuracy.	Limited congestion root-cause analysis.
Minovski et al. [11]	ML Framework	PRB Utilization, Active Users, CQI	Resource utilization strongly affected throughput degradation.	Complex feature engineering required.

1.4 Objectives

1.4.1 General Objective

The main objective of this thesis is to develop and evaluate a Random Forest-based machine learning model capable of learning the complex relationships between LTE Key Performance Indicators (KPIs) and throughput behavior in order to accurately predict LTE throughput congestion and identify the major KPIs contributing to the congestion.

1.4.2 Specific Objectives

The specific objectives of this thesis are:

- To review related works from the literature.
- Study LTE throughput and the KPIs that affect throughput performance and congestion.
- Collect KPI data from NMS specifically for BoleBulbula eNodeB cells
- Clean and prepare the dataset
- Study the correlations between different network parameters KPIs with throughput.
- Develop Random Forest model for throughput prediction.
- Evaluate the model's accuracy using standard metrics like RMSE, MAE, and R^2.
- Identify main KPIs driving congestion using the Random Forest model.
- Compare the model predictions with actual traffic conditions in BoleBulbula.

1.5 Methodology

The methodology used in this thesis focuses on developing a Random Forest Regression-based machine learning model to predict LTE throughput congestion. The first step of

the study is the collection of LTE KPI data from selected eNodeB cells located in the Bole Bulbula area of Addis Ababa. The required data is collected from Ethio Telecom's Network Management System (NMS).

The dataset includes different LTE network parameters related to throughput performance, signal quality, traffic conditions, and radio resource usage. Before applying the prediction model, the data is prepared using different preprocessing steps. These steps include managing missing values, removing invalid records, normalizing input features, and converting time information into a format suitable for analysis. Correlation analysis is also used to understand how different LTE KPIs are related to throughput changes.

The preprocessed dataset is then divided into training, validation, and testing sets. The training data is used to develop the Random Forest Regression model, while the validation data is used to adjust model parameters and improve performance. The final model evaluation is performed using the testing data with RMSE, MAE, and R^2 metrics.

In addition to predicting throughput, feature importance analysis is performed to identify the LTE KPIs that have more influence on congestion. This analysis helps to understand the main factors affecting network performance and provides useful information for LTE network optimization and improving QoS.

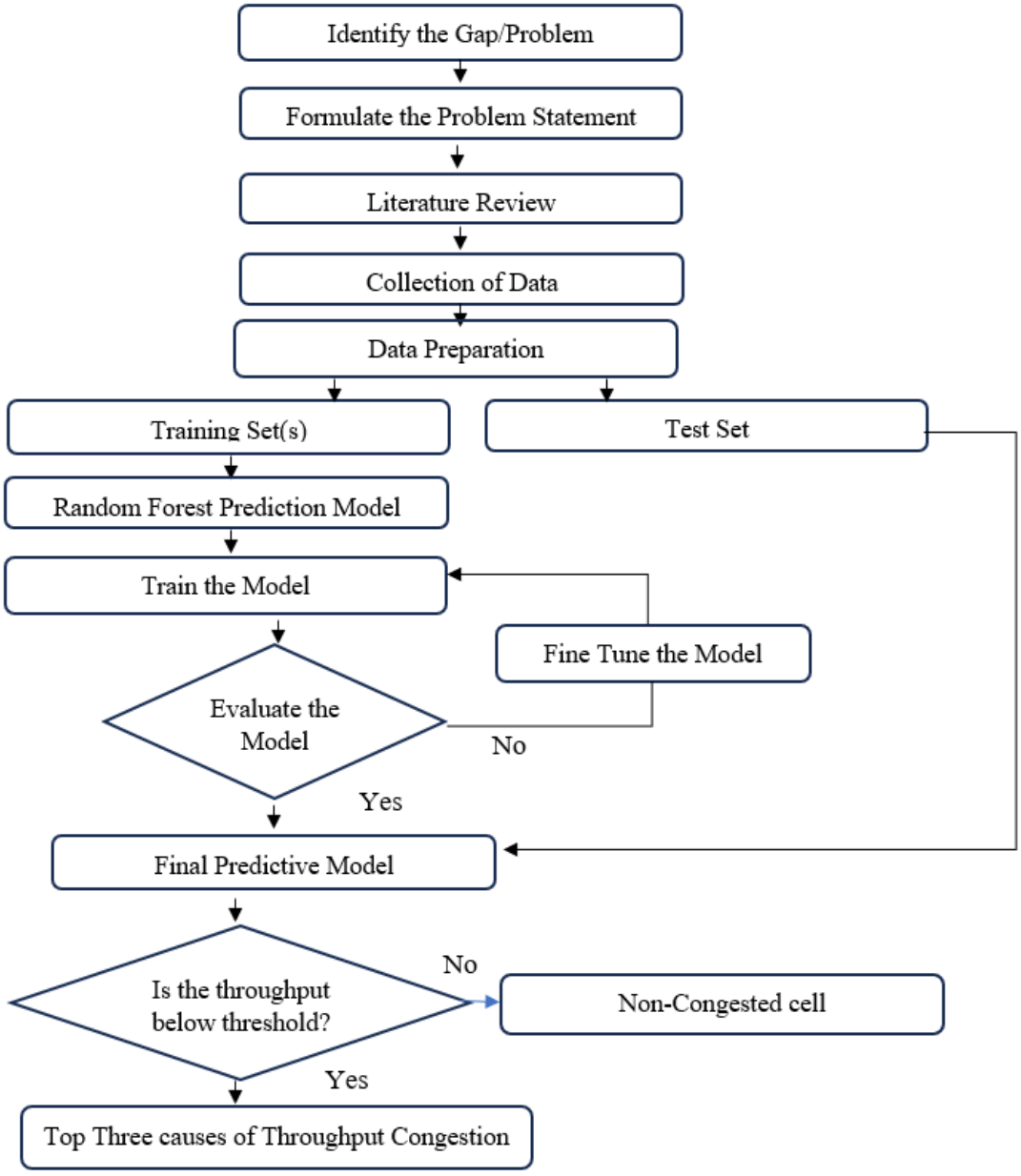


Figure 1.3: Methodology

1.6 Scope and Limitation

1.6.1 Scope of the Thesis

The scope of this thesis focuses on developing a machine learning model based on Random Forest Regression to predict LTE throughput congestion. The study specifically considers

4G/LTE network technology and utilizes LTE Key Performance Indicator (KPI) datasets collected from selected eNodeB cells in the Bole Bulbula area of Addis Ababa. The research focuses on analyzing LTE network parameters related to throughput performance, signal quality, resource utilization, and traffic behavior in order to predict throughput congestion conditions. In addition, the study investigates the importance of different LTE KPIs and identifies the major parameters contributing to network congestion using feature importance analysis. The work is limited to LTE network environments and does not include throughput analysis for 3G, CDMA, fixed broadband, or 5G standalone network architectures.

1.6.2 Limitation of the Thesis

This thesis is limited to historical LTE KPI datasets collected from selected eNodeB cells within a specific geographical area of Addis Ababa. The proposed model focuses only on LTE downlink throughput congestion prediction because downlink traffic carries the majority of mobile broadband data and experiences higher congestion conditions compared to uplink traffic within the studied LTE network environment. In contrast, uplink traffic showed relatively low congestion levels and lower resource utilization during the analysis period; therefore, uplink throughput prediction was not included in this study.

1.7 Outline of the Thesis

This thesis is organized into six chapters. Chapter One introduces the background of LTE throughput congestion, presents the problem statement, reviews related literature, and describes the objectives, methodology, scope, and organization of the thesis.

Chapter Two discusses LTE network architecture, LTE physical layer concepts, throughput measurement techniques, and the effects of congestion on network Quality of Service (QoS).

Chapter Three reviews machine learning approaches for LTE throughput prediction, including Random Forest, XGBoost, LSTM, and reinforcement learning methods, and ex-

plains the selection of the Random Forest model.

Chapter Four presents the proposed Random Forest-based prediction model, including data collection, preprocessing, normalization, feature encoding, correlation analysis, and model configuration.

Chapter Five presents the experimental results and performance evaluation of the proposed model, including throughput prediction accuracy, feature importance analysis, and congestion classification results.

Finally, Chapter Six summarizes the conclusions, provides recommendations, and suggests future research directions.

2 LTE Throughput

2.1 LTE Network Overview

The term Long-Term Evolution (LTE) is frequently used to describe the entire mobile network generation but it technically refers to the upgraded air interface introduced to advance 3G capabilities. This wireless interface establishes the critical connection between the subscriber's device termed User Equipment (UE) in 3GPP nomenclature and the carrier's infrastructure.The access side of this network is the Evolved Universal Terrestrial Radio Access Network (E-UTRAN), which replaced the 3G UMTS Radio Access Network (UTRAN) to deliver higher throughput and superior spectral efficiency. The network's core had to be completely redesigned to keep up with these faster data speeds which leads to Evolved Packet Core (EPC).

A main advantage of the EPS is its simplicity, by eliminating the unnecessary intermediary steps common in older networks, LTE creates a more direct route for data, resulting in faster and more efficient communication.[12]. Unlike its hierarchical predecessors, the LTE data plane minimizes intermediary nodes. This flattening trend actually began in 3GPP Release 7 with Internet-HSPA (I-HSPA), which started absorbing the responsibilities of the traditional Radio Network Controller (RNC) directly into the NodeB base station[13]. By eliminating the RNC node, the packet transmission path became shorter and more efficient than the multi-layered structures found in GSM and early UMTS releases.

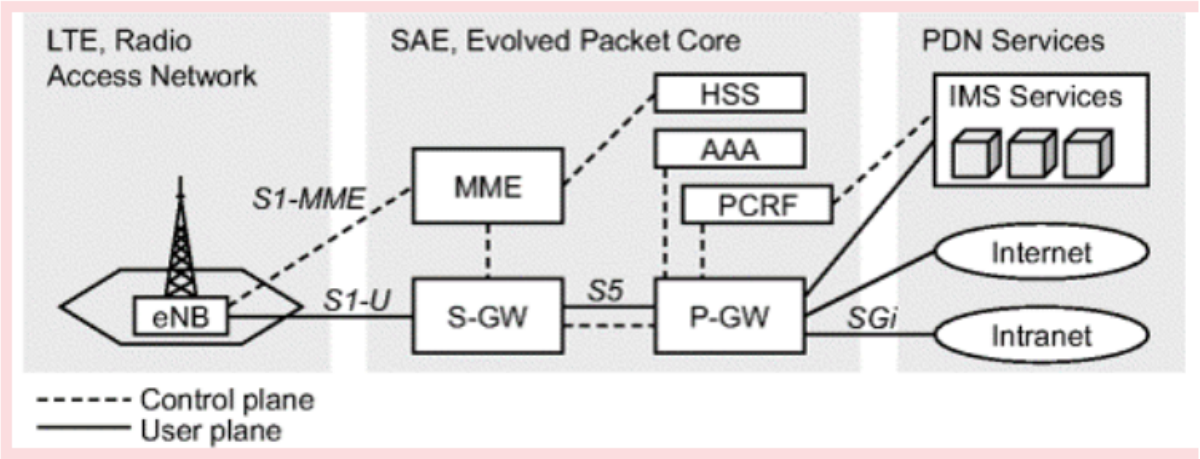


Figure 2.1: LTE Network Architecture[13]

2.1.1 LTE Radio Access Network

The E-UTRAN radio access network only consists of LTE base stations, which are called eNodeB or eNB (evolved NodeB) which are also the focus of thesis work. The eNodeB performs the primary functions of an LTE base station, including radio signal transmission and reception, modulation and demodulation, coding and decoding, as well as multiplexing and demultiplexing operations. It acts as the communication interface between the LTE network and the User Equipment (UE) while also handling important decision-making processes related to connection management, radio resource allocation, and traffic scheduling within the network.The eNodeB now includes basically all the Radio Resource Management functionalities which were previously concentrated on the additional RNC component, outside of the base stations, of the UTRAN system in 3G.

2.1.2 Evolved Packet Core

The Evolved Packet Core (EPC) is the framework that unified voice and data on a flat, Internet Protocol (IP)-based architecture for 4G LTE networks. Managed by the 3rd Generation Partnership Project (3GPP) starting in Release 8, the EPC was designed to optimize throughput, lower latency, and replace the separate circuit-switched (voice) and packet-switched (data) domains used in legacy 2G and 3G networks. Key Components of the EPC Architecture

The EPC is divided into two primary planes: the Control Plane (which handles signaling, security, and tracking) and the User Plane (which routes the actual data packets).Here are the components of the EPC [12]:

Mobility Management Entity (MME) [Control Plane]: responsible for device tracking, paging procedures, assigning bearer resources, and authenticating the user identity in conjunction with the HSS. It also manages handovers between eNodeBs.

Home Subscriber Server (HSS) [Control Plane]: A central database containing detailed subscriber-related and service profile information. It assists the MME in user authentication, access authorization, and roaming restrictions.

Serving Gateway (S-GW) [User Plane]: The local mobility anchor for the user data plane. The S-GW routes and forwards user data packets from the eNodeB into the core network while managing data paths during inter-eNodeB handovers.

Packet Data Network Gateway (PGW) [User Plane]: The interface between the internal EPC framework and external Packet Data Networks (PDNs), such as the public Internet or an operator's internal IP Multimedia Subsystem (IMS). The PGW handles IP address allocation for the device, Quality of Service (QoS) enforcement, and packet filtering/billing.

Policy and Charging Rules Function (PCRF) [Control Plane]: The node responsible for policy control decision-making and flow-based charging functionalities. It dictates how the PGW manages QoS and bandwidth for specific applications.

2.1.3 Packet Data Network

A Packet Data Network (PDN) is a digital communications network that breaks data down into small, routable units called packets to transmit them efficiently across shared channels. In mobile networks like LTE and 5G, a PDN represents an external data network such as the public Internet or a private corporate intranet that a subscriber connects to via an Access Point Name (APN) to access data services.

2.2 Physical Layer Aspects of LTE

The Physical Layer describes the physical transmission of bit sequences in LTE and the operation of the various channels according to the 3GPP specification [14]. The LTE physical layer encompasses several key aspects that enable high-speed and reliable wireless communication. LTE throughput is intricately tied to physical layer aspects, as these directly determine achievable data rates. Here below are some physical layer aspects of LTE.

2.2.1 Scalable Bandwidth and Single-Carrier Limits

In LTE , the physical layer operates strictly on a single-carrier framework where a User Equipment (UE) connects to only one frequency channel at a time. The system features a highly flexible, scalable bandwidth architecture supporting six distinct channel configurations: 1.4 MHz, 3 MHz, 5 MHz, 10 MHz, 15 MHz, and 20 MHz [15]. Because carrier aggregation does not exist in this baseline standard, the maximum transmission bandwidth allocated to a single user session is capped strictly at 20 MHz, which establishes a definitive hard ceiling for peak physical layer data rates.

2.2.2 Downlink and Uplink Multiple Access Schemes

The air interface is able to keep data transmissions well-organized while ensuring the device uses as little power as possible and this is possible by using two different types of multi-carriers. The downlink relies on Orthogonal Frequency Division Multiple Access (OFDMA), which divides the wideband channel into tightly spaced, orthogonal 15 kHz subcarriers to eliminate intra-cell interference and resist multi-path fading [16]. Conversely, the uplink implements Single Carrier Frequency Division Multiple Access (SC-FDMA) by applying a Discrete Fourier Transform (DFT) spread before subcarrier mapping, it mimics a single-carrier signal to drastically reduce the Peak-to-Average Power Ratio (PAPR), protecting mobile battery life. One of the main advantages of OFDM compared to single-carrier transmission schemes is its capability to operate efficiently under severe channel conditions. These conditions include high-frequency attenuation in long copper cables, narrowband interference, and frequency-selective fading caused by

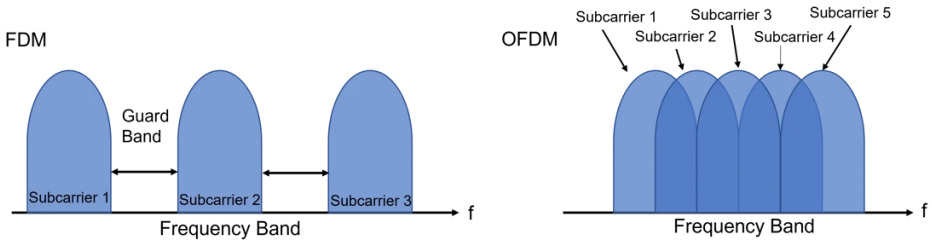


Figure 2.2: Orthogonal Frequency Division Multiplexing[17]

multipath propagation. OFDM mitigates these impairments without requiring complex equalization techniques.

2.2.3 Time-Frequency Resource Grid

All physical layer transmissions are organized within an immutable two-dimensional resource grid structured by time and frequency domains. Chronologically, the largest unit is a 10 ms radio frame divided into ten 1 ms subframes, with each subframe split into two 0.5 ms slots containing either six or seven OFDM symbols depending on the cyclic prefix length [18].The foundational building blocks of the Long-Term Evolution (LTE) air interface are structured hierarchically across the time and frequency domains. The absolute smallest physical unit is the Resource Element (RE), which spans a single subcarrier for the duration of one OFDM symbol. RE represents the atomic baseline of the radio grid, it is too granular for independent allocation. Instead, the network's Media Access Control (MAC) scheduler allocates data capacity in standardized dimensions known as Physical Resource Blocks (PRBs). Serving as the smallest schedulable unit of transmission that an eNodeB can assign to a User Equipment (UE), a standard PRB encompasses 12 consecutive subcarriers spaced at 15 kHz over a time duration of one slot (0.5 ms), totaling a bandwidth of 180 kHz.(180 kHz) over a single 0.5 ms time slot.

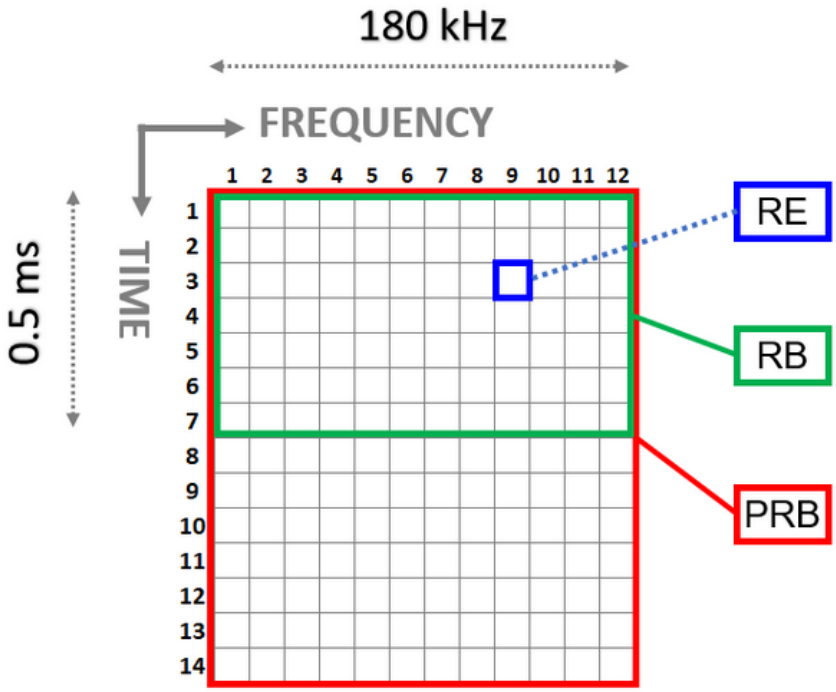


Figure 2.3: Physical Resource Block[19]

2.2.4 Baseline MIMO Antenna Configurations

Multi-antenna signal processing is natively integrated into baseline LTE to improve spectral efficiency and link reliability, though it remains restricted to lower-order matrix dimensions. For the downlink, the physical layer supports up to 2×2 or 4×4 MIMO configurations, enabling spatial multiplexing to transmit independent data streams to a single user or transmit diversity to strengthen cell-edge coverage . However, uplink spatial multiplexing is completely unsupported in this baseline era; the user device transmits only a single data stream via one power amplifier, relying strictly on receive diversity at the base station tower.

2.2.5 Adaptive Modulation and Coding

To extract the highest possible throughput out of a constantly fluctuating radio environment, the physical layer dynamically alters its transmission parameters based on real-time channel feedback from the mobile device. Based on the reported Channel Quality Indicator (CQI), the system selects between three modulation schemes: robust QPSK for

weak signals, 16QAM for mid-range conditions, and high-density 64QAM for optimal, low-noise environments close to the cell tower [20]. For error control and data correction over the air, the physical layer pairs these modulations with effective **Turbo Coding** for user transport blocks and convolutional coding for control signaling.

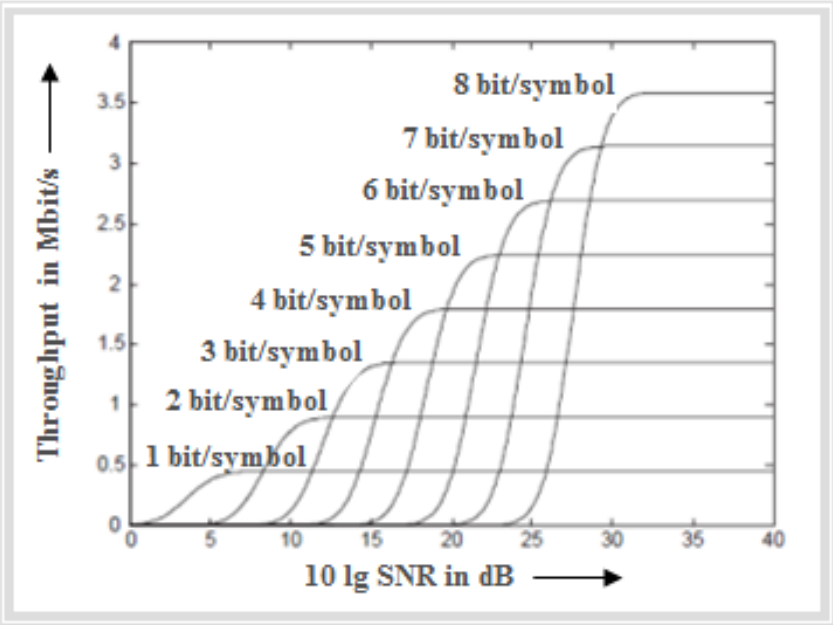


Figure 2.4: SNR Versus Throughput for Different Modulation Schemes[21]

The figure above shows the relationship between SNR and throughput for different modulation schemes. As channel quality improves, higher-order modulation techniques can be used, resulting in increased spectral efficiency and higher user throughput.

2.2.6 Physical Channel Mapping and Signaling

The physical layer acts as the bridge between software and hardware. It takes data and control signals from the MAC layer and maps them onto the actual physical channels needed for transmission. Downlink operations rely on the Physical Downlink Shared Channel (PDSCH) to carry user data, the Physical Downlink Control Channel (PDCCH) to manage resource scheduling assignments, and the Physical Broadcast Channel (PBCH) to deliver essential system initialization parameters. [16]. The uplink mirrors this setup using the Physical Uplink Shared Channel (PUSCH) for user uploads, the Physical Uplink Control Channel (PUCCH) for control acknowledgments, and the Physical Random Access Channel (PRACH) to manage initial network connection requests.

2.3 LTE Network Throughput Measurements

Network performance is the rate of successful data delivery through a communication channel. The amount of data traveling through the communication channel, the maximum bandwidth of the link, and the energy consumed in the channel are highly influential parameters of network performance. There are different ways to measure throughput in the LTE network.

2.3.1 Direct Measurements

User Equipment (UE) Measurements: These measurements are conducted directly on the user device. The actual data rates experienced by the user are measured, often using built-in tools or specialized applications on the device.Several commercial tools are also being developed to measure network throughput, including OOKLA and Iperf[6].

Signal Quality Indicators: Key parameters like Signal-to-Noise Ratio (SNR), Received Signal Strength Indicator (RSSI), and Modulation and Coding Scheme (MCS) are collected. These provide insights into the radio conditions that directly impact throughput. Network Measurements: Measurements are collected from the base stations (eNodeBs) within the LTE network.

eNodeB Measurements: Data on interference levels, resource utilization, and cell traffic are gathered. Performance Monitoring Tools: Network monitoring systems provided by the operator are used to collect data on critical performance indicators (KPIs) such as throughput, latency, and packet loss.

2.3.2 Indirect Measurements and Inference

Traffic Analysis and Network Management Systems (NMS)

LTE throughput can be analyzed using traffic measurements collected from Network Management Systems (NMS). Real-time and historical traffic measurements, utilization levels, and data volumes are collected to analyze sector and user throughput performance. This approach uses centralized Network Management Systems (NMS) and Performance

Management (PM) tools for network monitoring. The NMS collects counter data and PM statistics from network elements such as eNodeBs. These measurements are used to monitor traffic conditions, identify resource limitations, and estimate throughput performance without generating additional test traffic in the network.

Drive Testing

Drive testing uses test mobile units and diagnostic software installed in vehicles to collect radio frequency measurements such as RSRP, RSRQ, SINR, and application-level performance information while moving through selected geographical areas. These measurements provide information about radio conditions in specific locations and can be used to estimate throughput variations across different network areas.

Simulation-Based Modeling

Network simulators and software emulators are used to model network topology, scheduling behavior, and traffic variations in a controlled environment. This approach allows researchers to evaluate different network configurations and traffic scenarios before physical deployment. Simulation-based methods are useful for analyzing possible throughput changes and identifying potential network limitations.

2.3.3 Challenges in Throughput Measurement

Throughput measurement is challenging because throughput changes over time depending on user activity, location, and network conditions. Collecting detailed performance measurements can also introduce additional processing and measurement overhead. Furthermore, processing large KPI datasets requires appropriate analysis methods to identify useful performance information and understand network behavior.

2.4 Quality of Service Impacts of Network Load and Congestion

Throughput congestion occurs when the traffic demand in a cell exceeds the available radio resources, resulting in reduced user throughput and degraded Quality of Service (QoS). Under congested conditions, limited network resources must be shared among

multiple users, which can increase latency, reduce data rates, and affect overall user experience.

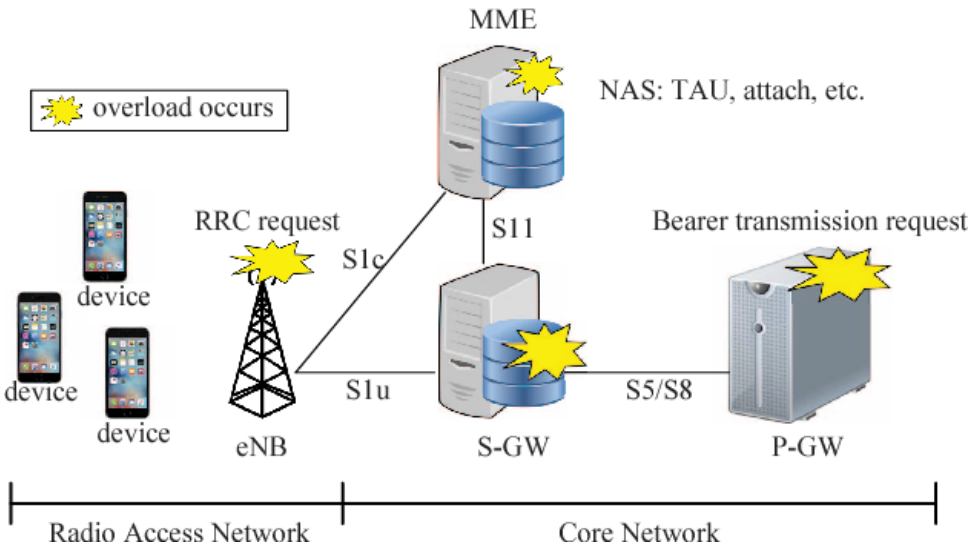


Figure 2.5: Throughput Congestion[22]

When LTE or LTE-A networks get congested, it causes a chain reaction that lowers both the network's performance and the user's experience.

2.4.1 Degradation of User-Plane Throughput

A common effect of network congestion is the reduction of user-plane throughput. When the total data demand from active users exceeds the available capacity of the Physical Resource Blocks (PRBs), the eNodeB Medium Access Control (MAC) scheduler distributes available radio resources among multiple User Equipments (UEs). As a result, each user receives fewer resources, which reduces the achievable data rate and increases the time required for data transmission.

2.4.2 Buffer Overflow and Packet Loss

When traffic load increases close to the processing and transmission limits of network nodes, packet queues at the eNodeB MAC layer may become overloaded. The increase in queue size results in higher buffer delay and increases the Round-Trip Time (RTT). When

the buffer reaches its maximum capacity, incoming packets may be discarded, causing packet loss [23]. For real-time applications, packet loss may require retransmission by upper-layer protocols, which can increase delay and further affect network performance [24].

2.4.3 Delay Variation and Packet Arrival Inconsistency

Network congestion can also affect the consistency of packet delivery. Changes in queue size cause variations in packet arrival times, resulting in packet delay variation, commonly known as jitter. For real time services like VoIP and video conferencing, even small amounts of jitter can cause noticeable lag, disrupting the flow of conversation or video quality[23]. When packets arrive outside the acceptable time range of the receiver buffer, they may be discarded, resulting in audio interruption and video quality degradation.

2.4.4 Application-Specific Performance Vulnerabilities

Different applications have different levels of sensitivity to congestion effects. Low-latency applications such as online gaming can experience performance degradation when delay increases. Multimedia applications using Adaptive Bitrate (ABR) streaming can reduce video quality or pause playback when available throughput decreases. In enterprise communication systems, frequent video quality reduction and buffering events can negatively affect communication performance.

2.4.5 Systemic Decline in User Satisfaction

The combined effects of reduced throughput, increased latency, and higher jitter can reduce the overall Quality of Experience (QoE). Continuous degradation of network performance may affect user satisfaction and reduce the ability of operators to meet required Quality of Service (QoS) levels. Therefore, maintaining acceptable throughput and delay performance is important for improving user experience and network reliability.

3 Machine Learning Models for LTE Throughput Prediction

This chapter reviews machine learning approaches related to LTE throughput prediction. The working principles of Random Forest, XGBoost, LSTM, Hybrid CNN-LSTM, and Reinforcement Learning are discussed to provide the background for selecting the proposed prediction model.

3.1 Machine Learning in Wireless Networks

LTE networks produce a large amount of KPI data that is difficult to analyze using traditional monitoring methods. Machine learning models help to analyze this data and find relationships between network conditions and performance. The increasing use of mobile broadband applications and connected devices has created more challenges for maintaining stable LTE performance and good Quality of Service (QoS). Traditional network optimization methods may not always respond effectively to the fast changes in wireless environments. Because of this, machine learning techniques are increasingly used in LTE and 5G networks for applications such as performance prediction, traffic analysis, and network optimization [25].

LTE networks continuously collect different measurements from eNodeB equipment through counters and KPIs. These measurements include throughput, Physical Resource Block (PRB) utilization, Channel Quality Indicator (CQI), latency, Block Error Rate (BLER), signal strength, and user movement information. These parameters provide information about the current condition of the network. Machine learning models can use historical KPI data to learn the relationship between network conditions and throughput changes. Compared with traditional mathematical models, machine learning methods can better handle complicated relationships between different network parameters [26].

Machine learning methods used in wireless networks are mainly divided into three groups: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning uses labeled historical data to develop prediction models for different tasks such as throughput prediction, traffic forecasting, handover improvement, and fault detection.

Regression methods such as Random Forest and XGBoost are commonly used to predict continuous network performance values, including LTE throughput [27]. Classification methods are also used to identify congestion conditions, service problems, and unusual network behavior.

Using machine learning in wireless networks provides many benefits, but there are still some challenges. These challenges include the need for large amounts of training data, high computational requirements, difficulty in explaining some models, and problems related to real-time implementation. Even with these limitations, machine learning continues to become an important technology for improving LTE networks and supporting future wireless communication systems.

Unsupervised learning is also useful in wireless network analysis, especially for clustering and detecting unusual network patterns. Since these methods do not require labeled data, they can group cells with similar traffic behavior and identify abnormal conditions that may indicate network problems. This is helpful for network planning because it can support the identification of high-traffic areas and better understanding of user distribution across LTE cells.

Deep learning methods are also used when dealing with large and time-dependent network data. Models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks can learn traffic patterns over time and help predict future throughput changes [28]. Convolutional Neural Networks (CNNs) can identify spatial relationships between network cells by analyzing patterns among neighboring cells. Combining CNN and LSTM models creates a hybrid approach that can capture both spatial and time-based information, making them useful for predicting congestion in areas with high network demand [29].

Reinforcement Learning (RL) has emerged as a powerful tool for day-to-day network optimization. By moving away from static configurations, RL enables an agent to learn through constant interaction with the network environment[30]. This creates a dynamic, responsive approach to managing spectrum allocation, load balancing, and congestion in real time.

Integrating these models into a live network environment presents practical hurdles, par-

ticularly regarding computational overhead, the demand for large training datasets, and the difficulty of ensuring model interpretability. However, despite these constraints, machine learning has become a foundational component in the evolution of LTE and the development of next-generation wireless communication systems.

3.2 Random Forest Regression

Random Forest Regression is an ensemble method that improves prediction accuracy by aggregating the results of multiple decision trees.This method works well for LTE KPI data since traffic patterns can be unpredictable, changing, and often nonlinear. By averaging these individual trees, the model effectively minimizes overfitting, stable and reliable way to predict network performance compared to single-tree methods.[31]. In LTE throughput prediction, Random Forest is suitable because network KPI datasets often contain noisy measurements, nonlinear relationships, fluctuating traffic conditions, and interference-related variations. This makes Random Forest suitable for LTE datasets containing variable traffic conditions and measurement variations.

Random Forest constructs multiple decision trees and combines their outputs to generate a final prediction with improves prediction consistency. It is also capable of processing large and non linear datasets with multiple KPI features such as Physical Resource Block (PRB) utilization, Channel Quality Indicator (CQI), Signal-to-Interference-plus-Noise Ratio (SINR), Block Error Rate (BLER), and user traffic load.

In LTE throughput prediction learning the relationship between network KPIs and throughput variation is important. By training on historical LTE network measurements, the model can estimate future throughput conditions under different traffic. Random Forest also provides feature importance analysis, allowing the identification of the KPIs that contribute to throughput degradation and congestion.

3.2.1 Decision Tree Ensemble Learning

Multiple decision trees are combined to improve prediction accuracy and reduce model variance. It works by repeatedly splitting data into smaller, organized groups based on

specific conditions until it reaches a final result. In this structure, the internal nodes serve as the decision points, while the leaf nodes represent the final predicted values at the end of each path. Each decision tree is simple and interpretable, also highly sensitive to noise

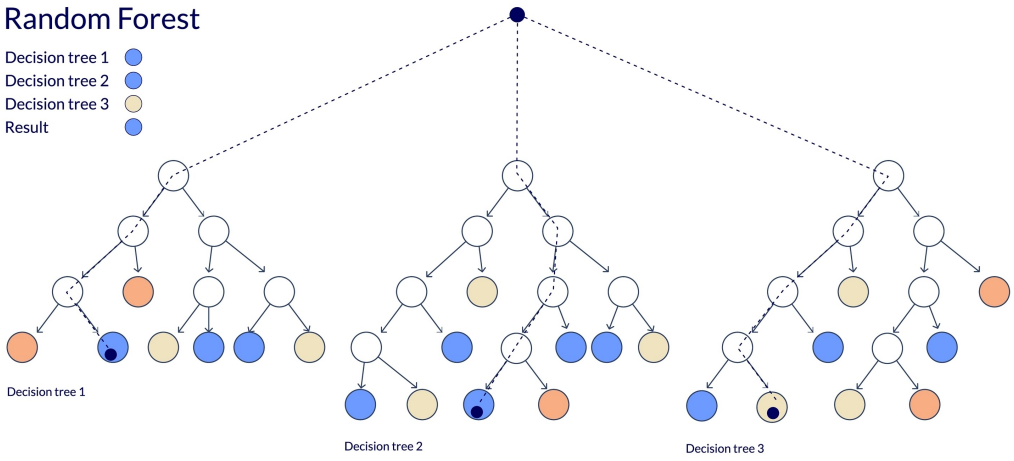


Figure 3.1: Decision Tree Ensemble in Random Forest Model[31]

and variations in training data. Small changes in LTE KPI measurements may cause a single decision tree to generate different predictions. RF overcomes this limitation by combining a large number of independently trained decision trees into a unified ensemble model[31]. This ensemble averaging process reduces prediction variance and improves robustness against overfitting. The mathematical representation of the Random Forest prediction is given by:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} f_b(x) \tag{3.1}$$

where:

- $\hat{y}$ represents the final predicted throughput value,
- B denotes the total number of decision trees,
- $f_b(x)$ represents the prediction generated by the b^{th} decision tree.

This ensemble design helps the model hold up well under dynamic conditions, ensuring performance stays steady even as LTE traffic patterns shift.

3.2.2 Bootstrap Aggregating

Bootstrap Aggregating (Bagging) is an important part algorithm that helps to create better predictions. Multiple bootstrap samples are randomly generated from the original training dataset. Each decision tree is trained independently using a different bootstrap sample. Instead of evaluating all available KPI features at every split point, the algorithm considers only a random subset of features. This randomness increases diversity among the trees and minimizes correlation between them.

3.2.3 Feature Importance Analysis

RF provides feature importance values that show the importance of each input variable to the prediction result. This is calculated by measuring how much each KPI contributes to reducing prediction uncertainty in all decision trees. Features that consistently improve node splitting and reduce prediction error receive higher importance scores. In Addition to forecasting throughput values, the model also provides insight into the underlying factors affecting LTE network performance.

3.3 XGBoost Regression

Extreme Gradient Boosting (XGBoost) is also an ensemble learning algorithm like Random Forest which base on gradient boosting. It builds decision trees sequentially to improve prediction by reducing previous errors[32] which can be used for regression and classification tasks.

Unlike Random Forest, which builds decision trees independently in parallel, XGBoost constructs trees sequentially which can improves accuracy by building trees one after another, where each new tree focuses specifically on fixing the mistakes made by the ones before it. It looks at the remaining errors known as residuals and learns how to adjust

for them. This step by step refinement allows the model to become increasingly precise over time.

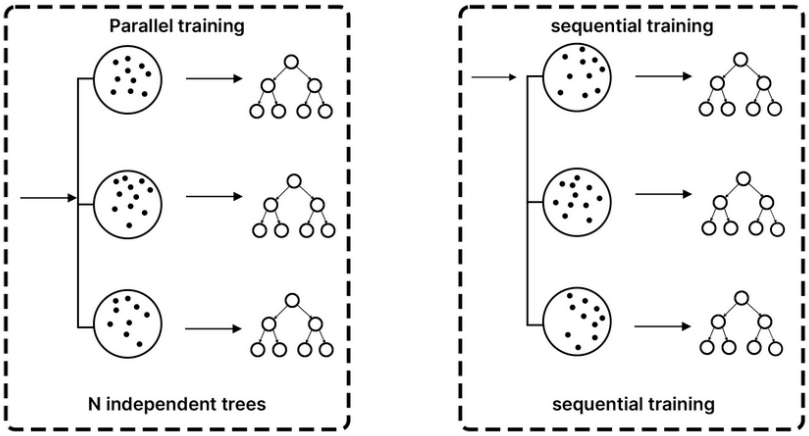


Figure 3.2: XGBoost vs. Random Forest[33]

XGBoost also incorporates several optimization techniques such as regularization, parallel computation, sparsity-aware learning, and efficient memory utilization. These features make the model computationally efficient and suitable for large-scale LTE KPI datasets.

Since XGBoost uses Gradient Boosting Mechanism which can refines its throughput predictions. By building a series of corrective trees for each iteration, it analyzes the difference between the actual and predicted values and corrects for those errors by training a new tree. This incremental approach allows the model to steadily minimize residuals and improve its predictive accuracy.

Mathematically, the XGBoost objective function is expressed as:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{3.2}$$

- $\mathcal{L}^{(t)}$ represents the objective function at iteration t,
- $l(y_i, \hat{y}_i)$ denotes the prediction loss function,
- $f_t(x_i)$ represents the newly added decision tree,
- $\Omega(f_t)$ represents the regularization term used to control model complexity.

In LTE throughput prediction, this boosting approach acts like a student constantly reviewing past mistakes to master complex patterns. It systematically learns how factors like PRB utilization, CQI, BLER, and SINR interact in non-linear ways, effectively mapping out how congestion and traffic behave. As each additional tree is added, the model sharpens its focus, becoming increasingly precise at forecasting how throughput will rise or fall under changing network conditions.

3.3.1 Regularization and Optimization

XGBoost uses built-in features to keep the model simple and stop it from over learning the training data. This helps the model stay accurate even when dealing with the noisy or messy data often found in network reports.when Readings are sometimes lost due to connection drops or errors. Instead of requiring someone to manually fill in these gaps, the system automatically figures out how to handle missing values while it builds the model, making it much more reliable for use. The model requires hyperparameter tuning to obtain suitable parameter settings and its sequential boosting structure can result in longer training time compared to simpler ensemble methods such as Random Forest.

3.4 Deep Learning Models for Throughput Prediction

Deep learning models are applied in wireless systems when network behavior depends on large datasets and sequential measurements. These models are effective at identifying the complex relationships and structural dependencies hidden in raw transmission data[34].
.

NMS in LTE networks generate KPI measurements continuously, which makes throughput prediction a time-dependent problem. Deep learning architectures such as Long Short-Term Memory (LSTM) and Hybrid CNN-LSTM models are capable of capturing both temporal and spatial dependencies within wireless network environments. These models can analyze time-dependent traffic patterns, user mobility effects, and congestion changes between LTE cells. this shows deep learning approaches have been applied in wireless network prediction problems.[29].

3.4.1 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) designed specifically to solve the vanishing gradient issue found in traditional RNN models [28]. Traditional RNNs have difficulty retaining information over long sequences, which limits their application in long-term traffic prediction. This limitations can be overcomed by using special memory cells and gates which act like a filter, deciding which information is important enough to keep from previous steps and which should be forgotten.

- **Forget Gate (f_t):** Controls which previous information should be removed from the memory state.
- **Input Gate (i_t):** Determines which new KPI information should be added to the memory cell.
- **Output Gate (o_t):** Controls the information passed from the memory cell to the output.

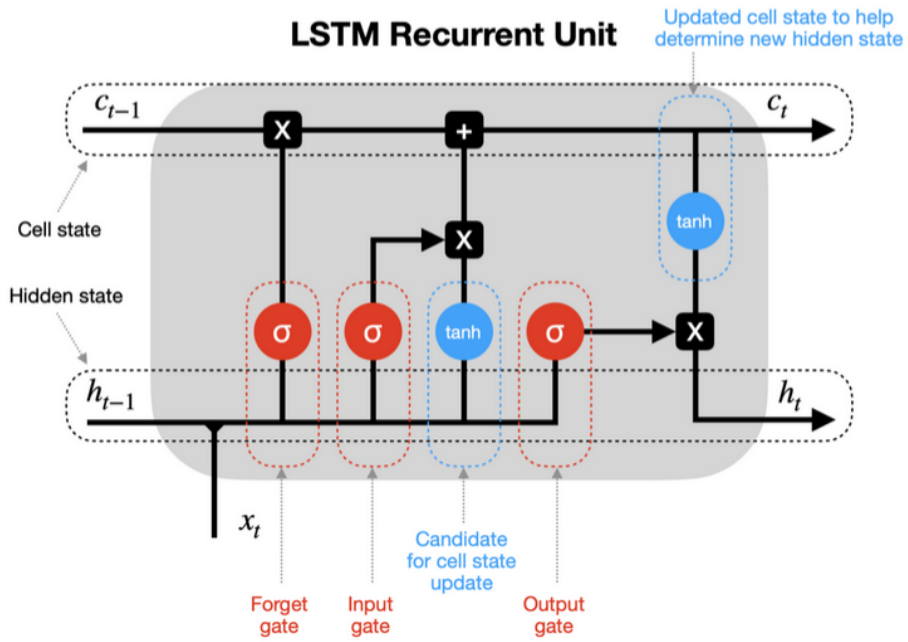


Figure 3.3: Long Short-Term Memory Neural Networks [35]

These gating mechanisms help the LSTM model maintain previous information and identify time-dependent LTE traffic patterns such as peak-hour congestion and throughput

variation. Therefore, LSTM models are suitable for throughput prediction problems involving sequential LTE measurements [35].

The main limitations of LSTM models are the requirement for large training datasets and higher computational resources.

3.4.2 Hybrid CNN-LSTM Architecture

The Hybrid CNN-LSTM combines Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to analyze both spatial and temporal characteristics of network data. This combination is useful for LTE throughput prediction because congestion patterns can vary with both time and network location.

In LTE systems, neighboring eNodeBs may influence each other due to user mobility, handover operations, interference, and overlapping coverage areas which means throughput degradation in one cell may affect nearby cells over time. The CNN-LSTM architecture combines spatial and temporal information that may not be captured by individual models.

The Convolutional Neural Network (CNN) component extracts spatial patterns from LTE network data by representing the network layout as a two-dimensional structure. Each element in the structure represents a network cell or eNodeB containing KPI measurements. Convolutional filters are then applied to identify local patterns within the input network data.

The mathematical operation of convolution is represented as:

$$(f * g)(n) = \sum_{m} f(m)\, g(n - m) \tag{3.3}$$

where:

- $(f * g)(n)$ represents the output feature map,
- $f(m)$ denotes the convolutional filter weights,

- $g(n-m)$ represents the input LTE KPI grid.

This combination of spatial and temporal analysis is excellent for tasks where both location and time matter. The downside for these models are harder to train, require significant computing power, and need very large datasets.

3.5 Reinforcement Learning Approaches

Reinforcement Learning is a machine learning approach where an agent learns actions through interaction with an environment using reward feedback.Unlike supervised learning techniques that rely on labeled datasets, Reinforcement Learning learns through a trial-and-error process by receiving rewards or penalties based on the actions performed[30].

The main advantage of RL is its flexibility. Because LTE networks are constantly changing due to users moving around, shifting signal quality, and unpredictable traffic—a system needs to adapt quickly. RL algorithms learn by constantly interacting with the network, adjusting their own decisions as conditions change without needing any intervention but using RL in a real-world LTE network isn't easy. It requires massive amounts of data and serious computing power. The learning process can also be quite slow, especially when the network is highly unstable.

In short, while RL is excellent for tasks like allocating resources or controlling the network, but it isn't usually the best tool for direct LTE throughput prediction.

3.6 Justification for Selecting Random Forest

Random Forest Regression was chosen for this study because it can handle many nonlinear LTE performance indicators KPIs in the data, and show which factors are most important in the predictions. LTE datasets are often complex sine the data contain noise, fluctuations, and changing traffic conditions, so a flexible model is needed to handel this.

The other reason for using Random Forest is its stability. It builds many trees and combines their results which can help to reduce errors and makes the predictions more reliable especially in changing network conditions.

Random Forest can also show the importance of different input features. This helps identify which LTE KPIs have the greatest impact on throughput, such as PRB utilization, CQI, BLER, and SINR. This makes the model useful not only for prediction but also for understanding the main causes of congestion.

Although models like XGBoost and deep learning can sometimes achieve higher accuracy, they usually require more complex tuning, larger datasets, and higher computing power. Random Forest offers a simpler and more practical balance between accuracy and interpretability.

For these reasons, Random Forest Regression was selected as the main model for LTE throughput prediction and congestion analysis in this work.

4 Proposed Random Forest Model and Data Preparation

This chapter details the Random Forest (RF) model used to predict throughput for selected eNodeB cells in the Bole Bulbula area. It outlines the data preparation steps taken prior to training and analysis. A key focus is the use of cyclical feature encoding for temporal data, a preprocessing technique implemented to improve the predictive accuracy of the Random Forest algorithm. Finally, the chapter summarizes the specific statistical metrics utilized for model evaluation.

4.0.1 Proposed Model Architecture and Hyperparameters

This thesis uses machine learning to handle two main tasks.The First one is to predict network performance. It forecasts how much throughput the network will handle over a specific time and determines if the network is congested or not congested in real time. This acts as a binary classification (where 0 means normal and 1 means congested) and the Second, it identifies the root causes of congestion, the model highlights the top three LTE metrics (KPIs) that are most responsible for any congestion detected. For Random Forest Regressor is a machine learning model in this study, 50 decision trees (*n_estimators*) were used. Each tree works like a small decision maker and looks at different parts of the network data to estimate the throughput value.

The model uses input features such as PRB utilization, BLER, and the sine/cosine time features. These values are given to all the decision trees, and each tree makes its own prediction based on different conditions in the data. Each tree follows a number of simple decisions using the network parameters to estimate the throughput.

Finally, the Random Forest combines the results from all the trees and calculates the average value to produce the final prediction, $\hat{y}$.

$$\hat{y} = \frac{1}{T} \sum_{i=1}^{T} f_i(x) \tag{4.1}$$

where $f_i(x)$ represents the prediction of the i-th decision tree. working as an ensemble smooths out random noise and reduces the impact of extreme data points, which results in much more stable and reliable predictions than a single tree could achieve.

To find the right size for the model, different values for the number of trees known as the $n_{\text{estimators}}$ parameter were tested. The goal was to find a balance between high accuracy and processing speed. While adding more trees usually makes a model more stable, there is a point of diminishing returns where the extra training time outweighs any small gain in performance. Based on the experimental results, the selected configuration is:

$$n_{\text{estimators}} = 50 \tag{4.2}$$

This number strikes the right balance, keeping the model fast and efficient without sacrificing the accuracy needed for LTE throughput and congestion analysis.

To ensure that the model remains generalized across the three-month dataset, the following architectural constraints were implemented:

- **Bootstrap Aggregating (Bagging):** Each tree is trained using a different random sample of the dataset, where some records can be selected more than once. This helps the Random Forest learn from different parts of the data and prevents the model from depending too much on specific periods, such as times with very high network congestion
- **Feature Sub-sampling:** During each split in the decision tree, the model randomly chooses a group of KPI features from the available data and uses them to decide the best way to split the data.(e.g., $\sqrt{total_features}$),This forcing the model to consider diverse KPIs simultaneously.
- **Mean Squared Error (MSE) Criterion:** The poposed model reduces the MSE to minimize the difference between the predicted and actual throughput values across all network.

4.1 Data Collection and Preparation

Data collection and preparation serve as the foundation for any reliable machine learning model used in LTE throughput prediction. High-quality throughput forecasting depends entirely on the accuracy and relevance of the underlying datasets. Proper preprocessing such as normalizing KPIs, handling missing values, and engineering temporal features ensures the model learns from consistent, clean signals rather than corrupted or erratic network noise. Without this rigorous preparation, even the most advanced algorithms will fail to capture the true operational dynamics of the network. In this study, LTE KPI data were collected from selected eNodeB cells in the Bole Bulbula area of Addis Ababa. The dataset contains different network performance measurements related to throughput, signal quality, resource usage, and traffic behavior. Before training the model, the data were carefully prepared by handling missing values, removing inconsistent records, normalizing the features, and applying cyclical feature encoding to time-related attributes. Eventhough massive data is collected only the most relevant KPIs affecting network congestion were selected to improve the prediction capability of the model. These data preparation steps help the Random Forest model learn the network behavior correctly and improve the accuracy of throughput prediction and congestion analysis.

4.1.1 Site Selection

In ethiotelecom's network of Addis Ababa, LTE generates the highest mobile data traffic, followed by 3G and 5G. Figure 4.1 below highlights the locations of congested LTE cells across different areas of Bolebulbulla.

This visualization clearly highlights where data demand is highest and where the network is most likely to hit bottlenecks. Considering this high data growth; this thesis focuses on 17 eNodeB sites in the Bole Bulbula area of Addis Ababa, which are prone to high demand and struggle to meet the throughput requirements of the subscribers.

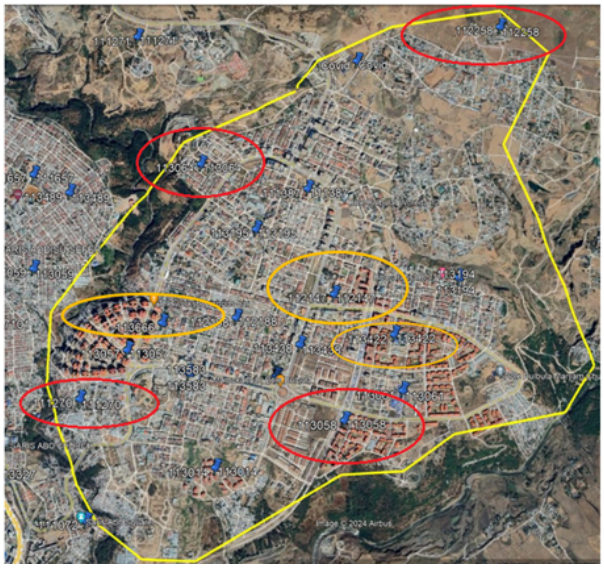


Figure 4.1: Sites with Higher Throughput at Bolebulbulla

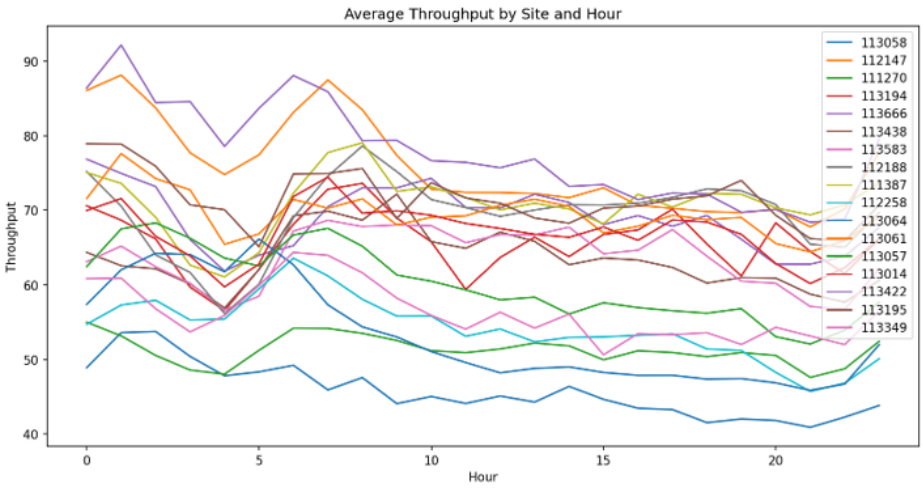


Figure 4.2: Average Throughput of The 17 Sites for July/2024 Data

4.1.2 Data Sources

The data for this thesis was collected from ethiotelecom's network management tool called NEC . This dataset contain 143 Key Performance Indicators (KPIs) extracted directly from the system for three consecutive months starting from June 2024. To identify highly correlated features from this data Kendall's Tau correlation method is used, a subset of these KPIs was selected as necessary features for throughput prediction

4.1.3 Data Cleaning

Data cleaning is crucial for ensuring data quality and reliability.Data cleaning make the data clean from noise and outliers which is the basic requirment for machine learning algorithm. The data cleaning process also includes the removal or imputation of missed values. To address missing values, there are two primary approaches:

Data Removal: If the proportion of missing values within the dataset is relatively low, removing data samples containing missing values can be a viable option[36].This approach helps maintain data consistency and prevents inaccurate model learning caused by incomplete samples. In this thesis, rows containing missing values were removed using a Python-based data preprocessing program. Since the amount of missing data in the collected LTE KPI dataset was limited, data removal was considered appropriate to improve dataset quality and ensure reliable model training for throughput congestion prediction.

Table 4.1: Eliminated Missing Values from Each Month's Data set

Month	Total Collected Data	Total Missed Data	Sites with Missed Data	Missed Data per Site	Missed Data (%)
June	133,714	567	112258	19	0.42
			113194	45	
			113195	171	
			113349	9	
			111387	324	
July	128,944	522	112255	54	0.40
			113014	153	
			113194	297	
			113349	18	
			111387	108	
August	105,184	297	113194	117	0.28
			113195	63	
			113583	9	

Missing Value Imputation: While not implemented in this study, imputation techniques, such as mean/median imputation or more sophisticated methods like K-Nearest Neighbors (KNN) imputation, can be employed to estimate and fill in missing values. This approach can help preserve valuable data that might otherwise be discarded.

4.1.4 Data Normalization

Normalization was applied because LTE KPI features have different value ranges, which can affect the learning process of the machiing model. By scaling the input features, the data becomes more suitable for model training and improves the stability of the prediction process.

The collected LTE KPI data contains values with different ranges, including some extreme values that may influence the model performance. Therefore, Robust Scaling was used before training the Random Forest model. This method uses the median and Interquartile Range (IQR) instead of the mean and standard deviation, making it more suitable for handling outliers. The scaling process reduces the effect of extreme KPI values and helps improve the reliability of throughput prediction[36].

$$x_{\text{scaled}} = \frac{x - \text{median}}{IQR} \tag{4.3}$$

where

$$IQR = Q_3 - Q_1 \tag{4.4}$$

where:

- Q_1 represents the 25th percentile,
- Q_3 represents the 75th percentile.

Normalization process was done using a Python program for this study. The `RobustScaler` function from the Scikit-learn library was applied to normalize the input features because the LTE KPI dataset contains outliers and different value ranges. [37]. Eventhough there are other scaling methods that use mean and standard deviation, RobustScaler uses the median and interquartile range to scale the data. This reduces the influence of extreme KPI values and allows the Random Forest regression model to perform more consistently during throughput congestion prediction.

For this study; a large number of parameters data collected from Ethio Telecom's network management system contains 144 KPIs. Due to this , it is important to select the features that have a strong relationship with throughput.The selected KPIs allow the model to learn the relationship between network conditions and throughput variation more effectively.

4.2 Cyclical Feature Encoding for Temporal Attributes

The amount of traffic in cellular networks and the changes in throughput usually follow some repeated patterns based on human activities. For example, network usage is usually higher during working hours and becomes lower during late night times. Time information is normally represented as a simple number from 0 to 23, but using this directly in a decision tree model can create a problem. The system may understand 0 and 23 as far apart values, even though they are actually very close in time because they represent midnight and the end of the same day.

To solve this problem, the time information was changed into a circular format. The 24-hour cycle was represented using sine and cosine functions, which helps the model understand that 23:00 and 00:00 are connected. This transformation keeps the repeating nature of time and allows the model to treat time as a continuous cycle instead of a normal numerical value.

In this study, cyclical feature encoding was applied to represent the time-based features in a better way. The original hour value was converted into two new features using sine and cosine functions. This conversion places the 24-hour period on a circular structure, where the beginning and end of the day are connected. The calculation of these cyclical features is shown below:

$$\text{Hour}_{\sin} = \sin\left(\frac{2\pi \times \text{Hour}}{24}\right) \tag{4.5}$$

$$\text{Hour}_{\cos} = \cos\left(\frac{2\pi \times \text{Hour}}{24}\right) \tag{4.6}$$

By utilizing this dual-coordinate feature transformation, the Random Forest algorithm is capable of accurately interpreting the continuous temporal dependencies and recurring congestion patterns across the dataset.

4.3 Correlation Analysis

Correlation analysis was performed to understand the relationship between LTE KPIs and throughput performance. A positive correlation indicates that two variables increase together, while a negative correlation shows that one variable increases as the other decreases.

The Pearson correlation coefficient is used to measure the linear relationship between variables. Its value ranges from -1 to 1 in which Values near 1 indicate a strong positive relationship and values near -1 indicate a strong negative relationship. A value close to 0 shows that there is no strong linear relationship between the variables. Although correlation analysis is useful for identifying relationships in data, it does not prove that one variable directly causes changes in another.

Pearson correlation was initially used to check the relationship between throughput and the LTE KPIs. During the analysis, the throughput values showed nonlinear behavior, which means Pearson correlation could not fully describe the relationship between the variables. For this reason, Kendall's Tau correlation coefficient was applied because it is more appropriate for analyzing relationships that do not follow a linear pattern.

```
Kendall Rank Correlation Coefficients (One-to-Many):
Column1 vs. Column2: 0.1119 (p-value: 0.1007)
Column1 vs. Column3: 0.6941 (p-value: 0.0000)
Column1 vs. Column4: 0.1480 (p-value: 0.0300)
Column1 vs. Column5: 0.0433 (p-value: 0.5255)
Column1 vs. Column6: 0.6524 (p-value: 0.0000)
Column1 vs. Column7: -0.0954 (p-value: 0.1616)
Column1 vs. Column8: 0.3333 (p-value: 0.0000)
Column1 vs. Column9: 0.3670 (p-value: 0.0000)
Column1 vs. Column10: 0.2195 (p-value: 0.0013)
Column1 vs. Column11: 0.1066 (p-value: 0.1181)
Column1 vs. Column12: 0.3395 (p-value: 0.0000)
Column1 vs. Column13: 0.1066 (p-value: 0.1181)
Column1 vs. Column14: 0.0965 (p-value: 0.1571)
Column1 vs. Column15: 0.1552 (p-value: 0.0228)
Column1 vs. Column16: 0.1493 (p-value: 0.0286)
Column1 vs. Column17: nan (p-value: nan)
Column1 vs. Column18: nan (p-value: nan)
Column1 vs. Column19: -0.0707 (p-value: 0.2998)
Column1 vs. Column20: 0.1360 (p-value: 0.0470)
Column1 vs. Column21: 0.1480 (p-value: 0.0300)
Column1 vs. Column22: 0.3524 (p-value: 0.0000)
Column1 vs. Column23: 0.3524 (p-value: 0.0000)
```

Figure 4.3: Kendall's Tau Correlation Output

Highly correlated labels can create redundancy and may cause overfitting. The p-value shows the statistical significance and helps gauge the chance of observing results by

random occurrence. Here are the features selected after correlation analysis:

1. **Integrity & Throughput (Target Group)**

   DL_User Throughput_Eng (Mbit/s): The primary target variable; it reflects the actual data speed experienced by an individual user.

   RO.Cell Downlink Average Throughput: Represents the aggregate data speed delivered by the entire cell.

2. **Accessibility KPIs**

   RO.RRC Connection Success Rate (%): The probability that a mobile device successfully initiates a connection with the eNodeB.

   E-RAB Setup Success Rate (%): The probability of establishing a data bearer after the initial connection is made.

   Cell Availability (%): The percentage of time the cell was operational and providing service.

3. **Retainability & Reliability**

   E-RAB Drop Rate_Eng (%): Frequency of unexpected session terminations.

   DL BLER (%): Downlink Block Error Rate; measures data packet corruption due to channel noise.

   Reliability Index: A composite metric calculated as $100 - \text{DL BLER} - \text{RRC Congestion}$.

4. **Mobility KPIs**

   RO.Intra-frequency HO Success Rate (%): The success rate of user handovers between cells on the same frequency.

5. **Load & Utilization**

   DL PRB Utilization (%): Indicates the usage level of the physical resource blocks.

   CCE Utilization_Eng (%): Measures the signaling load on the control channel.

   RRC Congestion Ratio (%): The rate at which connection requests are blocked due to capacity limits.

6. **Traffic & User Demand**

   TrafficVolume (GB): The total payload volume downloaded during the interval.

L.Traffic.ActiveUser.DL.Avg: The average number of users simultaneously active in the downlink.

7. Coverage & Interference

RO.Cell Minimum Interference: Represents the baseline signal cleanliness within the cell.

TA (<1KM) / TA (>1.95KM) (%): Proportions of users located very close to or far from the tower.

Net Proximity: A geometric distribution index calculated from the Timing Advance (TA) differences.

KPI	3GPP Standard	Description	Value Range	Relation with Throughput	ET Threshold
PRB	TS 36.211	Smallest time-frequency resource unit used for LTE downlink transmission.	0–100	Higher PRB utilization increases data carrying capacity but may also increase congestion during heavy traffic conditions.	70%
CQI	TS 36.213	Metric reported by the UE indicating radio channel quality between the UE and eNodeB.	0–15	Higher CQI enables higher modulation and coding schemes, improving throughput performance.	–
CCE	TS 36.211	Control Channel Elements allocated for LTE control signaling information.	–	Higher CCE allocation for control signaling reduces available resources for user data transmission, potentially lowering throughput.	10%
E-RAB Setup Success Rate	TS 36.300	Represents the successful establishment of radio access bearers between UE and the core network.	–	Poor E-RAB performance may indirectly affect throughput through inefficient resource allocation and signaling overhead.	10%
BLER	–	Ratio of incorrectly received data blocks to the total transmitted blocks.	–	High BLER increases retransmissions and negatively affects throughput performance.	10%
RSRP	TS 36.214	Average received power of LTE reference signals.	-140 dBm to -44 dBm	Higher RSRP values indicate stronger signal conditions and support higher throughput performance.	–
RSRQ	TS 36.214	Quality metric combining signal strength and interference measurements.	-19.5 dBm to -3 dBm	Poor RSRQ values indicate high interference and degraded throughput performance.	–
SINR	–	Signal-to-Interference-plus-Noise Ratio used to measure signal quality.	–	Higher SINR improves modulation efficiency, reduces transmission errors, and increases throughput.	–

4.4 Model Configuration

The dataset was divided into three subsets: training (70%), validation (15%), and testing (15%). The training dataset was used to train the Random Forest model, while the validation dataset was used to adjust the model parameters and improve the performance. After that, the testing dataset was used to check how well the model works on new data that it did not see before.

The training data contained more than 310,000 records, which helped the model learn the relationship between LTE KPIs and downlink throughput changes. Before training the model, the Quality Index feature was removed because it was calculated using information related to the target output. If this feature was kept, it could cause data leakage and make the model performance look better than the actual result.

This research aims to produce highly accurate throughput predictions (Mbit/s) for the downlink throughput, which are then verified against actual field data using standard metrics like the Mean Absolute Error (MAE) and the coefficient of determination (R^2). In addition to throughput prediction, feature importance analysis was applied to determine the LTE KPIs that contributed most to throughput variation. This analysis helps identify the parameters associated with congestion conditions in the selected cells.

Table 4.2: Summary of Proposed Random Forest Model Parameters

Parameter Category	**Model Specification**
Algorithm Selection	Random Forest Regressor
Ensemble Configuration	100 Decision Trees (n_estimators)
Data Partitioning	70% Training / 15% Validation Set/ 15% Testing
Target Variable (y)	DL_User Throughput_Eng (Mbit/s)
Feature Dimensions (x)	23 (6 Time Features + 17 Network KPIs)

The time-series data was divided based on the order of time to keep the relationship between previous and future data points. After splitting the data, the target classes were checked to make sure that both congested and non-congested conditions were included properly in the training, validation, and testing datasets with a balanced distribution.

5 Results and Discussion

In this chapter, the proposed Random Forest regression model prediction accuracy and performance evaluation are discussed briefly using independent test dataset. To assess the advantage of using multiple network metrics, the model performance is compared using a full set of multivariate KPIs against those relying solely on simple, univariate throughput data.Finally, the Random Forest feature importance distributions are analyzed to diagnose the primary root-cause KPIs driving network throughput degradation.

5.1 Throughput Forecasting Analysis

By combining the 17 network KPIs with temporal data—specifically hour, day, and month the model functions as a throughput forecaster. It analyzes historical trends and the current state of the selected eNodeB cells to determine existing capacity and usage baselines. Once these patterns are established, the model extrapolates these relationships to predict throughput conditions for the subsequent time interval.

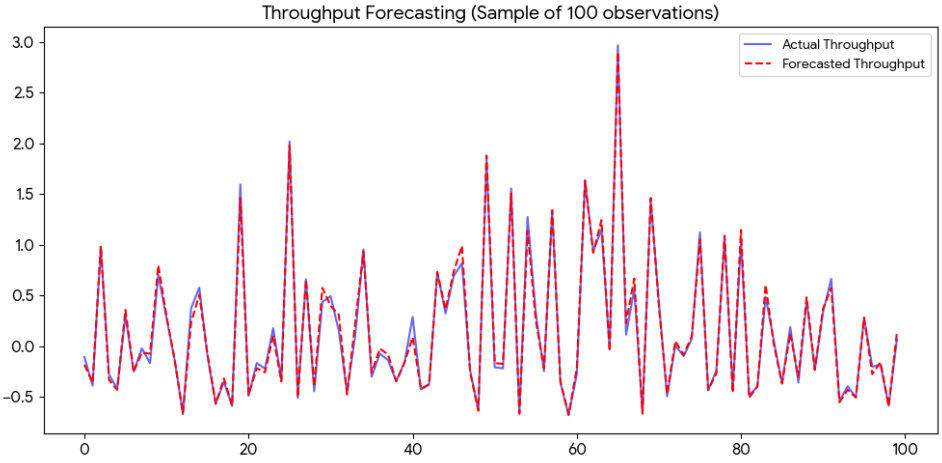


Figure 5.1: Model Throughput Forecast Test

The scatter plot below illustrates the high correlation between the actual data measurements collected form ethiotelecom's NMS and the predicted throughput from the model. The tight clustering of data points along the red dashed diagonal line shows; the predicted values show close agreement with the measured throughput values across different network conditions. Whether the network is struggling with low-speed congestion or running at peak capacity, the model's predictions remain consistent. The result indicates that the Random Forest model can capture variations in LTE throughput under different operating conditions.

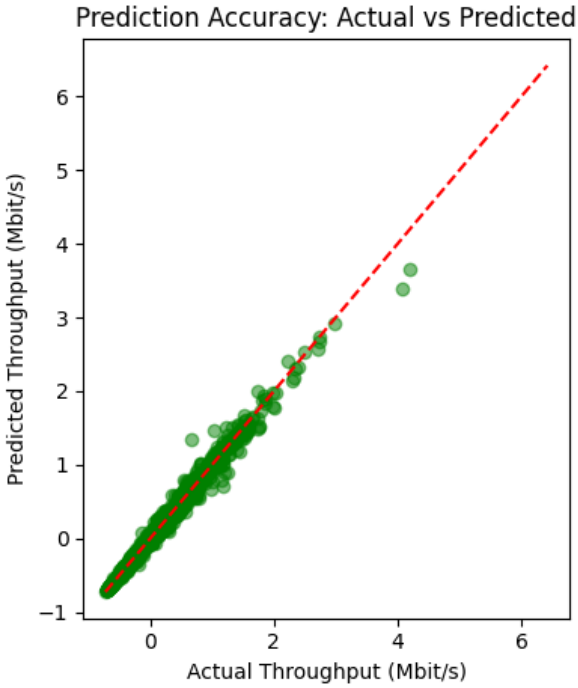


Figure 5.2: Actual vs. Predicted Throughput Correlation

5.2 Feature Importance Analysis

Feature importance analysis was used to identify which input variables had the highest contribution to throughput prediction. The results indicate that Physical Resource Block (PRB) utilization and average cell throughput are the most significant predictors of user experience. This suggests that, within the local network environment, traffic volume and resource consumption are the key factors determining whether a connection maintains high quality or experiences congestion.

The Cell Downlink Avg Throughput acts as a reference point that helps you distinguish between normal operation and congestion, and it accounted for 46.1% of the importance. The second baseline parameter is DL PRB Utilization, which represents the resource load on the physical channels and had a 43.2% impact on the prediction. The third most important factor was the average number of active users, which contributes 10.3% of importance and represents the number of simultaneous users sharing the same cell resources. Other factors, such as traffic volume, CCE utilization, and time-of-day cycles, were also vital to the model. They provided the necessary context to capture how the network shifts from heavy usage during peak hours to the quiet lulls observed late at nigh

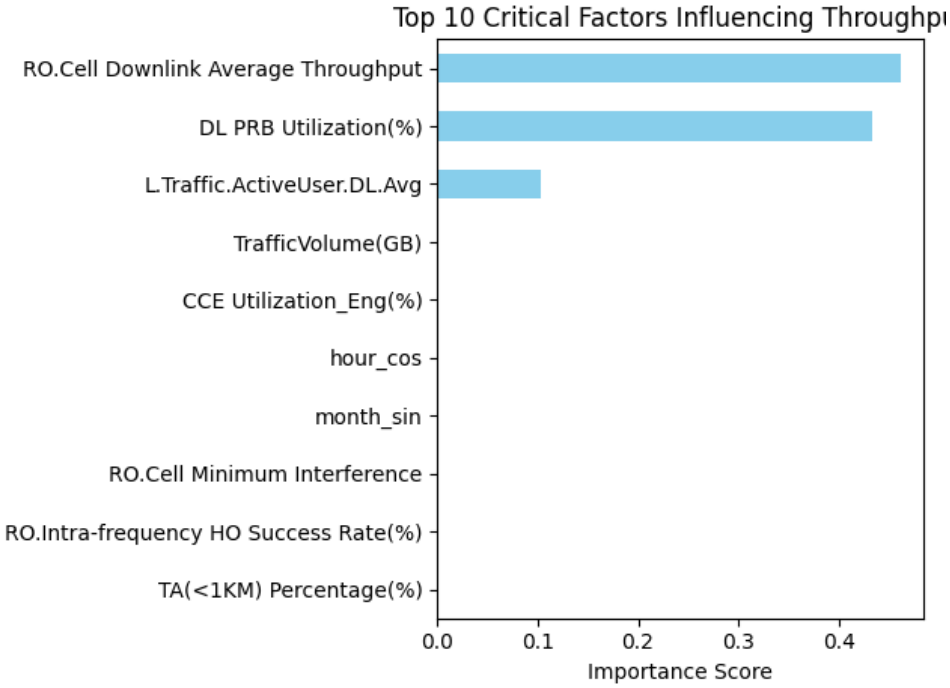


Figure 5.3: Top 10 Feature Importance for Throughput Prediction

Table 5.1: Feature Importance Analysis for Congestion Prediction

Feature	**Importance (%)**	**Description**
Cell Downlink Avg Throughput	46.1	Basic measurement of the available capacity of a cell.
DL PRB Utilization	43.2	Resource load on physical channels.
Avg. Number of Active Users	10.3	Simultaneous users sharing cell resources.
Other Factors	0.4	Traffic Volume, CCE Utilization, and Cyclical Time.

5.3 Residual and Error Distribution

To verify that the model is unbiased, we need to look at how its prediction errors are distributed. The sharp peak tightly centered around zero is a great sign that for most test cases, the model's predictions are very close to reality. In addition to this accuracy to predict, the symmetry of this distribution on figure 5.4 shows that the model isn't leaning toward consistently over-predicting or under-predicting throughput, which can add point on the reliability of the model.

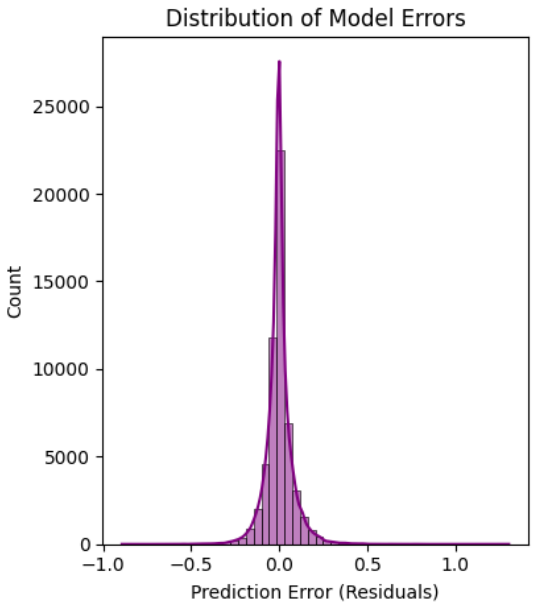


Figure 5.4: Prediction Error Distribution

5.4 Target Variable Correlation Heatmap

The chart below shows which features have the strongest statistical linear relationship with throughput. While the Random Forest model captures complex non-linear patterns, this chart illustrates the raw strength of each KPI's connection to the target variable.

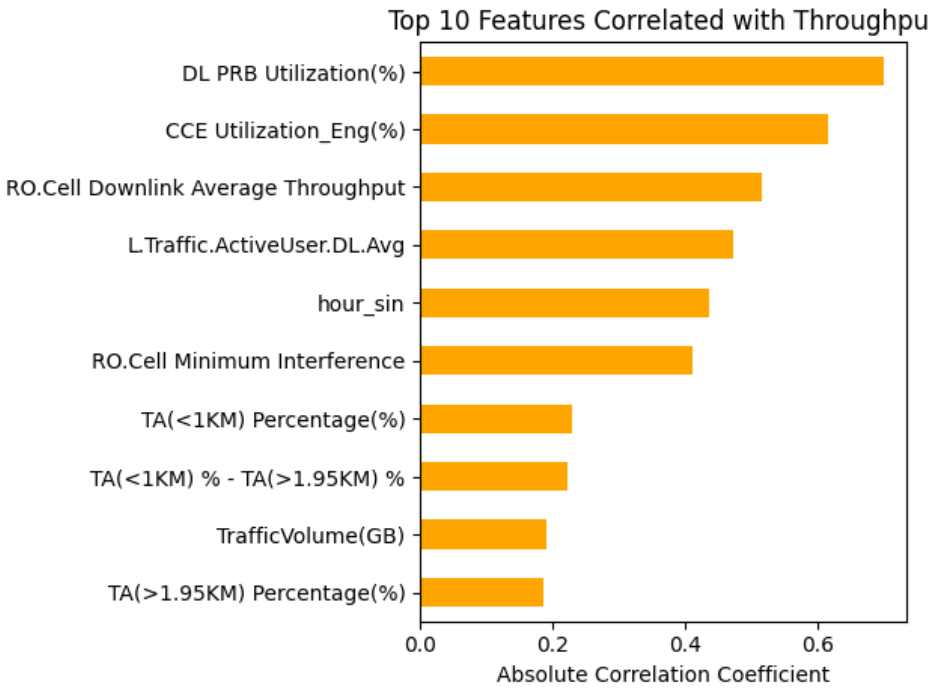


Figure 5.5: KPI Linear Correlation Heatmap with Target Throughput

5.5 Binary Congestion Classification Performance

In ethiotelecom 70th percentile of PRB utilization is considered as a threshold to consider if a cell is congested or not congested. With this in mind, the model is trained to classify the network state into a simple binary: 0 for healthy operation or 1 for congestion.

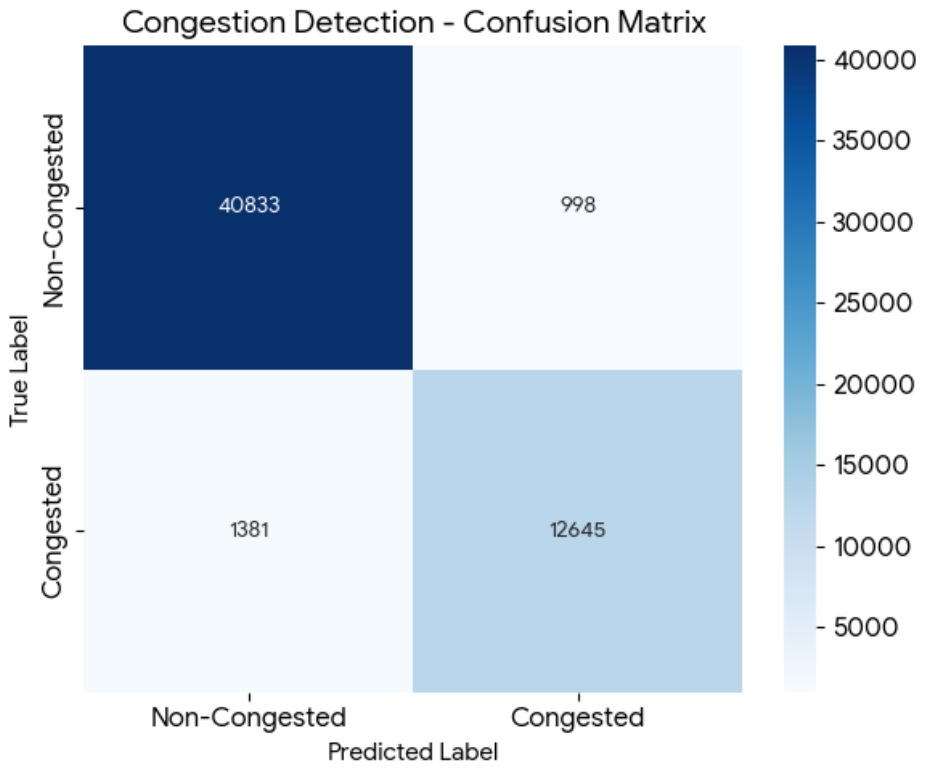


Figure 5.6: Binary Congestion State Prediction Output

Table 5.2: Performance Evaluation for Congestion Detection Model

Metric	**Calculation Formula**	**Result**
Accuracy	$\frac{TP+TN}{N} = \frac{12645+40833}{55857}$	**0.9574**
Precision	$\frac{TP}{TP+FP} = \frac{12645}{12645+998}$	**0.9268**
Recall	$\frac{TP}{TP+FN} = \frac{12645}{12645+1381}$	**0.9015**
F1-Score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	**0.9140**

Where: TP = True Positive; TN = True Negative; FP = False Positive; FN = False Negative; N = Total Samples ($TP + TN + FP + FN$).

5.6 Model Evaluation Metrics and Performance Analysis

The Random Forest model achieved an R^2 value of 0.987, indicating that the model explains 98.7% of the variation in downlink throughput. Furthermore, the MAE value of 0.050 Mbit/s and RMSE value of 0.079 Mbit/s demonstrate that the model maintains a low level of prediction error on the test dataset. This shows that the model remains steady and reliable, even when handling the heavy traffic spikes seen during peak hours.

Table 5.3: Model Accuracy

Metric	**Score**	**Engineering Interpretation**
R^2 **Score**	0.989	This indicates an exceptional fit, where the model successfully captures 98.7% of the variance influencing throughput fluctuations.
MAE	0.043	On average, the model's predictions deviate by only 0.043 Mbit/s, showing it is reliable for precise capacity planning.
RMSE	0.071	This low value confirms that the model rarely makes large errors, ensuring consistent performance even during peak traffic hours.

6 Conclusion and Recommendations

6.1 Conclusion

This thesis developed and evaluated a Random Forest Regression model to predict downlink throughput and congestion in LTE cells in the Bole Bulbula area. I also looked at how different LTE Key Performance Indicators (KPIs) relate to throughput changes under real traffic conditions. To handle the daily ups and downs in network usage, I added cyclical time features using sine and cosine transformations. This small addition helped the model understand traffic patterns much better.

The final model performed great, achieving an R^2 value of 0.987. This means it could explain almost 99% of the variation in throughput. checking the feature importance using Mean Decrease in Impurity (MDI), Downlink PRB Utilization came out as the strongest factor by a clear margin. CCE Utilization, BLER, and Timing Advance also showed noticeable influence.

One unexpected observation was just how dominant PRB Utilization was. During data collection from Ethio Telecom's NMS, which implies that many cells in Bole Bulbula frequently crossed 70–80% PRB usage during peak hours, which directly led to slower speeds for users. This confirms that resource exhaustion is the main cause of congestion.

In summary, machine learning particularly RF proved to be a practical and useful tool for predicting LTE congestion using real network KPI data. The model not only gives good predictions but also helps engineers see the real reasons behind performance problems.

6.2 Recommendations

The results show that PRB Utilization has the biggest impact on throughput congestion. Therefore, Ethio Telecom should give priority to capacity improvement in areas like Bole Bulbula. Techniques such as Carrier Aggregation (CA) and better optimization of MIMO configurations could help ease the pressure on the existing resources.

Another important point is that looking at KPIs one by one doesn't always give the full

picture. I recommend developing a combined performance indicator that brings together throughput, PRB utilization, and BLER. This would make it easier for engineers to quickly judge the overall health of a cell.

The prediction model can also support a shift from the current reactive approach to a more proactive one. If implemented in the Network Management System, it could alert engineers about possible congestion before users start complaining about slow internet.

6.3 Future Work

Although the model gave good results,in the future, it would be useful to test the same approach on 5G New Radio (NR) data to see how it performs with newer technologies like beamforming and different numerology options.

It would also be interesting to add more mobility features, such as handover events, user movement patterns, and geographical information. Since Bole Bulbula has a mix of residential and commercial users, understanding mobility better could improve predictions during peak commuting times.

Finally, expanding the dataset to include more sites across Addis Ababa and incorporating real-time KPI streams would make the model much more practical for daily network operations at Ethio Telecom.

References

[1] Ethio Telecom, *Mobile network traffic and KPI analysis.* Ethio Telecom, Addis Ababa, Ethiopia, May. 25 2024. Internal network performance database.

[2] M. Malekzadeh and A. A. A. Ghani, "3-sector cell vs. omnicell: Cell sectorization impact on the performance of side-by-side unlicensed lte and 802.11 ac air interfaces," *IEEE Access*, vol. 7, pp. 122315–122329, 2019.

[3] Ethio Telecom, *Ethio Telecom OSS Data.* Ethio Telecom, Addis Ababa, Ethiopia, Aug. 20 2024. Internal network performance database.

[4] N. Stepanov, D. Alekseeva, A. Ometov, and E. S. Lohan, "Applying machine learning to lte traffic prediction: Comparison of bagging, random forest, and svm," in *2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pp. 119–123, IEEE, 2020.

[5] L. Alho, "Machine learning based mobile network throughput classification," in *IEEE Wireless Communications and Networking conference*, pp. 1–110, 3GPP, 2021.

[6] S. Abolfazli, Z. Sanaei, S. Y. Wong, A. Tabassi, and S. Rosen, "Throughput measurement in 4g wireless data networks: Performance evaluation and validation," in *2015 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, pp. 27–32, IEEE, 2015.

[7] E. Eyceyurt, Y. Egi, and J. Zec, "Machine-learning-based uplink throughput prediction from physical layer measurements," *Electronics*, vol. 11, no. 8, p. 1227, 2022.

[8] L. Alho, A. Burian, J. Helenius, and J. Pajarinen, "Machine learning based mobile network throughput classification," in *2021 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, IEEE, 2021.

[9] X. Wang, Z. Wang, K. Yang, Z. Song, J. Feng, L. Zhu, and C. Deng, "Deep learning based traffic prediction in mobile network-a survey," *Authorea Preprints*, 2023.

[10] D. Verma, H. Saraf, and S. H. Gupta, "Prediction of user throughput from network parameters of lte network using machine learning," *Mobile Networks and Applications*, vol. 28, no. 1, pp. 244–253, 2023.

[11] D. Minovski, N. Ogren, C. Ahlund, and K. Mitra, "Throughput prediction using machine learning in lte and 5g networks," *IEEE Transactions on Mobile Computing*, 2021.

[12] 3GPP, "General packet radio service (gprs) enhancements for evolved universal terrestrial radio access network (e-utran) access," Technical Specification TS 23.401, 3rd Generation Partnership Project (3GPP), 2024.

[13] S. Ahmadi, *LTE-Advanced: A Practical Guide to the Next Generation of Wireless Communications*. Oxford, UK: Academic Press, 2014.

[14] J. Zyren and W. McCoy, "Overview of the 3gpp long term evolution physical layer," *Freescale Semiconductor, Inc., white paper*, vol. 7, pp. 2–22, 2007.

[15] S. Sesia, I. Toufik, and M. Baker, *LTE – The UMTS Long Term Evolution: From Theory to Practice*. Chichester, West Sussex, UK: John Wiley & Sons, Ltd, 2nd ed., 2011.

[16] E. Dahlman, S. Parkvall, and J. Sköld, *4G: LTE/LTE-Advanced for Mobile Broadband*. Oxford, UK: Academic Press, 2nd ed., 2013.

[17] G. L. Stuber, J. R. Barry, S. W. Mclaughlin, Y. Li, T. G. Pratt, and A. Colodny, "Orthogonal frequency division multiplexing for wireless communications," *IEEE Communications Surveys & Tutorials*, vol. 6, no. 3, pp. 2–14, 2004.

[18] 3GPP, "Policy and charging control architecture," Technical Specification (TS) 23.203, 3rd Generation Partnership Project (3GPP), 2020.

[19] TeleTopix, "Resource block and physical resource block in lte," 2024. Accessed: 2026-05-25.

[20] 3GPP, "Evolved universal terrestrial radio access (e-utra) and evolved universal terrestrial radio access network (e-utran); overall description; stage 2," Technical Specification TS 36.300, 3rd Generation Partnership Project (3GPP), 2023. Standard reference for LTE Load Balancing and Handover procedures.

[21] J. G. Proakis, *Bandwidth-Efficient Modulations*. McGraw-Hill, 4th ed., 2001.

[22] X. Dong, W. Fan, and J. Gu, "Predicting lte throughput using traffic time series," *ZTE Communications*, 2015. Accessed: 2026-05-25.

[23] Y. Bandung and I. Sumardi, "A methodology for characterizing real-time multimedia quality of service in limited bandwidth network," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 14, no. 4, pp. 1534–1543, 2016.

[24] B. Nguyen, N. Choi, M. Thottan, and J. Van der Merwe, "Simeca: Sdn-based iot mobile edge cloud architecture," in *2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, pp. 503–509, IEEE, 2017.

[25] S. H. Alsamhi, F. A. Almalki, O. Ma, M. S. Ansari, and S. K. Gupta, "Machine learning techniques for wireless networks: A review," *Sensors*, vol. 21, no. 18, p. 6100, 2021.

[26] R. Li, Z. Zhao, Q. Sun, I. Chih-Lin, C. Yang, X. Chen, M. Zhao, and H. Zhang, "Intelligent 5g: When cellular networks meet artificial intelligence," *IEEE Wireless Communications*, vol. 24, no. 5, pp. 175–183, 2017.

[27] K. Wang and Y. Zhang, "Lte network traffic prediction based on machine learning," *IEEE Access*, vol. 5, pp. 15833–15841, 2017.

[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[29] C. Zhang, P. Patras, and H. Haddadi, "Deep learning based mobile traffic prediction," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2191–2215, 2019.

[30] H. Mao, R. Netravali, and M. Alizadeh, "Neural adaptive video streaming with pensieve," 2017. Excellent reference for RL in congestion/throughput control.

[31] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[32] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[33] A. Vidhya, "Getting started with xgboost: A beginner-friendly tutorial," 2023. Accessed: 2026-05-25.

[34] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. Chapter 9 provides the mathematical definition of convolutional operations.

[35] MathWorks, "Long short-term memory networks." MathWorks Documentation Center, Deep Learning Toolbox, 2026. Accessed: May 2026.

[36] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2011.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.