4G DATA ANALYSIS AND DOWNLINK THROUGHPUT PREDICTIVE MODELLING AROUND THE GEOGRAPHIC REGION OF GAZIPUR AND DHAKA

by

Shafi Muhtasim Chowdhury (190021326)
Safwan Islam (190021219)
Rafrafin Islam (190021223)

Department of Electrical and Electronic Engineering
Islamic University of Technology (IUT)
Gazipur, Bangladesh

June 2024

4G DATA ANALYSIS AND DOWNLINK THROUGHPUT PREDICTIVE MODELLING AROUND THE GEOGRAPHIC REGION OF GAZIPUR AND DHAKA

Approved by:

Md. Samiur Rahman
Supervisor and [Associate/Assistant] Professor
Department of Electrical and Electronic
Engineering Islamic University of Technology
(IUT)
Board Bazar, Gazipur-1704.

Date: ......

Table of Contents

List of Tables

List of Figures

List of Acronyms

3GPP	Third Generation Partnership Project
ACF	Auto Correlation Function
ARIMA	AutoRegressive Integrated Moving Average
CNN	Convolutional Neural Network
COs	Course
HSPA	High Speed Packet Access
HSDPA	High Speed Downlink Packet Access
IEEE	Institute of Electrical and Electronics Engineers
IoT	Internet of Things
ITU	International Telecommunication Union
KNN	K-Nearest Neighbor
KPI	Key Performance Indicator
LTE	Long Term Evolution
MLE	Maximum Likelihood Estimation
OBE	Outcome Based Education
PACF	Partial Auto Correlation Function
QoS	Quality of Service
QXDM	Qualcomm Extensible Diagnostic Monitor
RF	Random Forest
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
RSSI	Received Signal Strength Indicator
SNR	Signal-to-Noise Ratio

Abstract

This project presents a comprehensive study on optimizing 4G data analysis and downlink throughput modelling, integrating technical rigor with societal considerations. As 4G networks continue to serve as a backbone for modern telecommunications, enhancing their efficiency and performance is crucial for meeting growing data demands and supporting emerging applications. This research aims to develop advanced models and algorithms that improve network performance while addressing broader impacts on safety and societal aspects.

A systematic, theory-based approach is employed, grounded in the fundamentals of telecommunications engineering. Literature review encompasses seminal works and recent advancements in 4G technology, signal processing, and data analysis, providing a robust theoretical foundation. Important concepts in machine learning such as data preprocessing and data augmentation are explored in depth and utilized heavily in this project. Additionally, industry standards and protocols from bodies like the IEEE, 3GPP, and ITU are reviewed to ensure the research aligns with current best practices.

The methodology involves the use of diverse resources, including datasets that were collected by hand from selected regions of interests in Gazipur and Dhaka, specialized software tools for model simulation, data analysis and network performance data collection, and case studies of real-world 4G network deployments. Advanced statistical methods, machine learning algorithms, and optimization techniques are employed to analyze data and model throughput. The project tries to emphasize the importance of data anonymization, security, and compliance with data protection regulations to address ethical and privacy concerns.

Innovative approaches are central to this project. Novelty in the project arises from the fact that the dataset used in the project in its entirety was collected by using an application named G-Nettrack. The dataset contains network performance data, that were localized to selected regions around Gazipur as well as other regions in Dhaka such as Uttara, Banani, Mirpur, Dhanmondi, Baridhara, Rampura, Mohammadpur and Bashundhara. Machine learning algorithms and models are developed were chosen based on their ability to work best on the dataset, and were used to optimize

downlink throughput, leveraging machine learning, and AI techniques for predictive analytics and performance enhancement.

The results demonstrate that machine learning models can simulate and predict network downlink throughput with acceptable standards of accuracy, and validated through rigorous analysis and evaluation. Detailed data analysis reveals patterns and trends that inform the optimization models, while comparative analysis with existing studies highlights the advancements achieved. The research also addresses conflicting requirements, such as security with accessibility, and attempts to propose solutions that achieve optimal trade-offs.

By providing a holistic view that encompasses technical excellence and societal responsibility, this project attempts to help set a foundation for ongoing research and development in the field of telecommunications. It highlights the importance of innovation, collaboration, and ethical considerations in advancing 4G networks to meet the evolving needs of society.

Chapter 1: Introduction

The advent of 4G brought about a significant transformation in the telecommunications sector, providing users with unparalleled data speeds and greatly enhancing the overall user experience. The introduction of 4G brought about significant changes not only in personal and business communication but also expanded the possibilities for various applications such as online gaming, video streaming, and connected autonomous vehicles. As the world progresses further into the era of digital technology, there is a growing demand for high-speed data that surpasses the currently available levels. The growing need for optimizing network performance and enhancing quality of service (QoS) has been emphasized by this rising demand. The purpose of the project is to thoroughly analyze 4G network performance data to identify patterns, trends, and potential network challenges. The ultimate objective is to develop predictive models that can be used to optimize networks in areas of interest around Gazipur and Dhaka.

1.1. Background and Motivation

In the last ten years, improvements in cellular network technology have significantly transformed the methods of connecting and communicating with the world. The shift from 3G to 4G technology has introduced a wide range of services and capabilities that were previously inconceivable. 4G technology enables a diverse array of essential applications in modern life by providing faster data speeds and reduced latency. Smartphone users are becoming more dependent on their devices not just for communication, but also for various activities like work, entertainment, social networking, and accessing real-time information. The growing dependence on this technology has resulted in a substantial surge in internet traffic, placing continuous strain on cellular networks to satisfy user needs. Cellular network operators must address the challenge of maintaining network performance and QoS while accommodating the growing demand for data for the increasing population of network users in Bangladesh. Conventional methods of managing networks and distributing resources are inadequate for dealing with the dynamic and intricate network applications of today. As a result, there is an increasing demand for creative solutions that can effectively tackle these difficulties (Figure 1.1).

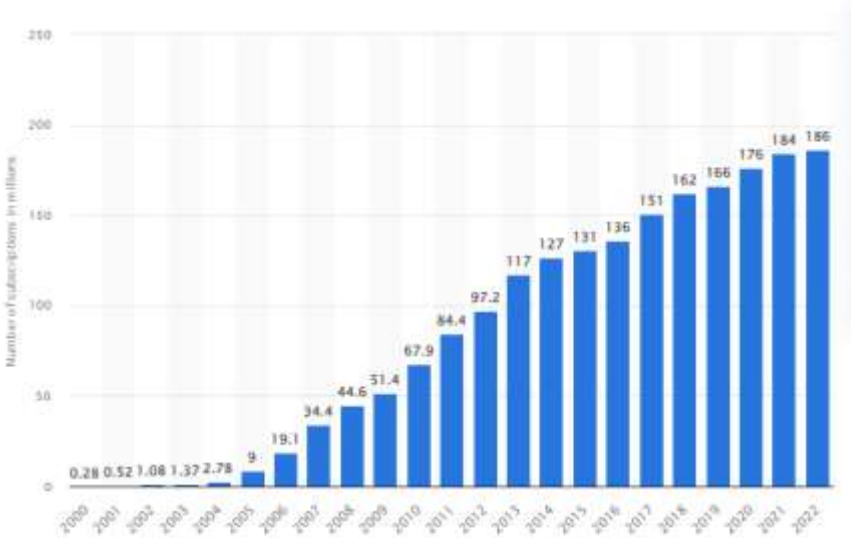

Figure 1.1 : Number of cellular subscriptions in Bangladesh from 2000 to 2022

The introduction of 4G technology has significantly worked towards enhancing data speeds and connectivity. With the increasing reliance on mobile data for personal and professional use, the demand for robust and reliable 4G networks has surged. This is particularly relevant in rapidly developing regions such as Gazipur and around my areas around Dhaka, which are all urban centers in Bangladesh that have witnessed exponential growth in population and digital infrastructure. Gazipur, a key industrial hub, and Dhaka, the capital city and economic heart of Bangladesh, represent critical areas where efficient 4G networks are essential for sustaining economic activities, improving quality of life, and supporting various smart city initiatives. The dynamic urban environment and diverse population of these cities pose unique challenges and opportunities for telecommunications providers. Network optimization in such densely populated and industrially active regions requires a sophisticated understanding of usage patterns, signal propagation, and potential bottlenecks.

One effective method to enhance network QoS and tackle scalability problems is to implement a predictive approach for resource allocation and network management. The primary concept underlying this approach is to predict in advance the occurrence of fluctuations in network connectivity and utilize these predictions to devise proactive measures. By doing so, the network can promptly allocate additional resources to the user, ensuring a smooth and uninterrupted experience. The process of dynamic allocation is becoming increasingly crucial due to the growing prevalence of bandwidth-intensive applications, the increasing exposure to media and consumption of social media, and the widespread use of connected and autonomous vehicles, which are responsible for this trend.

This work is motivated by the need to overcome the challenges faced by 4G networks. It aims to optimize network performance by utilizing predictive models. By making precise forecasts about network conditions, it is possible to develop tactics to guarantee reliable QoS and improve the overall user experience. This proactive approach is not only beneficial for ongoing experiments but also crucial for future advancements in telecommunications.

1.2. Literature Review

The evolution from 1G to 4G mobile communication technology has fundamentally altered how data is transmitted and received. 4G network development has increased demand for effective data analysis and predictive modelling that would enhance network performance, especially in highly populated areas such as Gazipur and Dhaka.

Many researchers have studied and proposed potential answers to the problem of predictive modelling and throughput prediction over a number of years. Abdullah bin Shams investigated how variations in user mobility affected downlink resource scheduling and consequently how it affected the downlink throughput [1]. He concluded that resource scheduling utilizing the round robin scheduling algorithm at high user equipment speeds allowed for seamless network connections and as a result better overall network scheduling and throughput.

A study conducted by Fakirah ira et al. looked through how LTE network performance parameters varied in selected areas around Indonesia [2]. They utilized a software named G-NetTrack to collect network performance data around regions of interest and showed how each of the important parameters varied. Similarly, El-Saleh et al. utilized the G-Nettrack software to collect network performance data of five different mobile network operators in Malaysia and made a comprehensive discussion relating to how networking conditions of the different network operators varied [3]. The investigation conducted by Ekeocha, Elechi and Nosiri also studied how the key performance indicators (KPIs), namely the RSSI, RSRQ and RSRP, were affected across selected regions in Port Harcourt, Nigeria [4]. Elsherbiny et al. collected datasets along public transit bus trips in Kingston, Ontario at specific time intervals [5]. The final dataset was used to generate 4G downlink throughput predictive models, that were evaluated using different metrics to ensure high model accuracy and performance. They found that Random Forest (RF) algorithm achieved the best performance among the other machine learning algorithms used in the investigation [6].

Kamaris and Nickerson put forward the idea of using a connectivity map for throughput estimation [7]. They studied the relationship between the received signal strength and the throughput in Wi-Fi networks and were able to find that dynamic variations in the network conditions led to short average lifetimes of connectivity map predictions. Similarly, Pögel and Wolf proposed the idea of using connectivity maps for predicting different network parameters such as the RSSI, latency and bandwidth in a vehicular context [8]. They performed many drive tests, which are simply the process of collecting required data while driving and collected the necessary information relating to network performance parameters in a HSDPA network. Following this, they continued their work to help to enhance user experience through better handover between different network technologies, adaptive video streaming, etc. by using their previously collected dataset [9].

The PROTEUS system interface makes it possible to make accurate and instantaneous predictions regarding throughput, which was made possible by the work of Xu et al. [10]. An analysis of the performance of the network over the course of the previous twenty seconds is carried out, and regression trees are utilized to generate forecasts. This analysis is performed to generate forecasts. Through the utilization of this method, the significance of utilizing data that is up to date to make accurate predictions for the short term is brought to light. For predicting the throughput of mobile networks, Liu et al. conducted an analysis and comparison of seven different algorithms [11]. The data that they used was trace-driven and came from 3G/HSPA networks. Data on throughput that occurred during periods of inactivity was used by their models to make predictions regarding future throughput, which occurs every 300 seconds. These predictions were made using the data that was collected. It has been demonstrated that it is possible to make use of a variety of machine learning techniques to forecast the performance of a network.

The issues relating to the application of throughput prediction algorithms in 4G LTE networks were investigated by Jin in 2015 [12]. They obtained from the QXDM toolkit to arrive at an estimate of the rollout speeds of 4G LTE networks, and then later developed LinkForest, a machine learning based framework for throughput prediction. The framework takes in low-layer information such as RSRP, RSRQ, SNR, etc. to predict instantaneous link bandwidth. Samba et al. also performed throughput predictions using the RF algorithm [13]. They were able to improve their approximations by incorporating additional data that was provided by network operators. This data included the average throughput of cells as well as connection speeds. This all-encompassing approach exemplifies the process of integrating different kinds of data to maximize the performance of the model and achieve the best possible results. Table 1.1 below illustrates the summary findings of the literature review.

Table 1.1: Summary of literature review

Sl. No	Author	Focused on	Location	Tools Data Source
1.	Abdullah bin Shams et al.	Impact of user mobility on downlink resource scheduling	Bangladesh	HetNet
2.	H. Elsherbiny	4G Predictive models	Kingston, Ontario	Python
3.	Ekeocha et al.	Performance evaluations of KPIs for LTE	Port Harcourt, Nigeria	Drive tests(DTs), Python
4.	El-Saleh et al.	Evaluation of network performance metrics of MNOs	Malaysia	G-NetTrack Pro
5.	Fakirah ira et al.	LTE Signal Analysis	Indonesia	G-NetTrack Pro
6.	Kamakris and Nickerson	Connectivity map for throughput estimation in Wi-Fi	New Jersey, USA	Netstumbler
7.	Pogel and Wolf	Connectivity map for predicting different network parameters	Braunschweig, Germany	HSPDA network data
8.	Pogel and	Optimization of	Braunschweig,	Ns-2 simulator

	Wolf	vehicular applications	Germany	
9.	Xu et al.	PROTEUS for instantaneous throughput prediction	Michigan, USA	Regression trees, 20 seconds of observed data, PROTEUS
10.	Liu et al.	Comparison of algorithms for mobile network throughput prediction	Hong Kong	3G/HSPA network data, stationary scenario, 300-second intervals
11.	Ruofan Jin	LinkForecast for 4G LTE throughput prediction	Connecticut, USA	QXDM toolset, machine learning-based framework
12.	Samba et al.	Throughput prediction using random forest algorithm	Lannion, France	Crowdsourcing data, additional data from network operators

1.3. Problem Analysis

While 4G has advanced far in recent years, it still faces substantial issues in maintaining optimal network performance and consistent QoS due to the rapid increase in community traffic. Since it depends on multiple elements, forecasting downlink throughput—a critical indicator for directing proactive network capabilities—is primarily multifaceted. The components include the time of day, geographical location, geographic proximity, and achieved signal strength. Predicting these changes with enough accuracy will be the main challenge to allocate resources in advance and guarantee continuous consumer feedback.

Conventional methods of managing communities frequently fall short in addressing these problems since they are impulsive. They typically respond to community problems only after they have already happened, which results in less than ideal performance and worse user experiences. There is a fundamental necessity for a more actively involved strategy that can detect possible network problems in advance and resolve them before they affect users.

The fact that community environments are diverse contributes to making the problem more detrimental. Urban community management presents special issues because of the dense population, different physical factors, and irregular usage patterns. Predicting network performance is a complex process that involves factors including individual movement, weather patterns, and residential buildings. Therefore, for effective community optimization, models that can take into account all of these factors and generate credible forecasts must be developed.

1.4. Objectives

The primary objective of this challenge is to tackle the urgent challenges encountered by 4G networks because of continuously growing demand for high-speed data. This assignment aims to enhance community performance and QoS by utilizing advanced information analysis and predictive modelling techniques. The mission's objectives are to identify crucial factors, create predictive models, offer optimization strategies, and provide actionable insights for related concerns. Following is the list of specific objectives of this project:

- *Identification of key factors influencing 4G network performance:* Perform a comprehensive analysis to identify the multiple factors that have a significant impact on the performance of 4G networks. This involves analyzing factors such as signal intensity, geographic coordinates, topography, temporal variations, and user movement behaviors.
- *Development of predictive models for early detection of potential network issues:* Create advanced predictive models capable of forecasting potential network issues before they occur. These models will utilize machine learning techniques and time series forecasting methods to predict future network conditions accurately.
- *Proposal of effective optimization strategies to enhance 4G QoS:* Create and suggest efficient strategies for enhancing the performance of 4G networks. These strategies will prioritize enhancing resource allocation, optimizing load balancing, and effectively managing the network to guarantee users receive high-quality service.

- *Insights for network operators, policymakers, and researchers to improve overall 4G network efficiency:* Provide valuable insights and recommendations for network operators, policymakers, and researchers. These insights will help in making informed decisions to enhance the efficiency and effectiveness of 4G networks.
- *Development of predictive models that incorporate socioeconomic factors for enhanced accuracy:* Incorporate socioeconomic factors into the predictive models to enhance their precision. This will require analyzing factors such as population density, economic activity, and user behavior patterns in various regions.

The goal of this project effort is to enhance network performance, optimize resource allocation, and ultimately improve the quality of service (QoS) for all 4G users in Bangladesh. The proactive and predictive approach will help enable the development of network architectures that are more resilient and efficient, capable of effectively meeting the increasing demands of the digital age. The results of this project will not only be advantageous for existing 4G networks but will also yield valuable knowledge and approaches that can be utilized in future iterations of mobile networks, such as 5G and beyond.

Chapter 2: Overview of the Project

The dynamic and rapidly changing telecommunications industry requires creative strategies to enhance network efficiency and optimization, particularly in 4G networks. These networks face increasing demands for high-speed data and consistent service quality, making it imperative to prioritize their development. The system is structured into significant procedures, each of which plays a role in enhancing network connectivity.

At first, the main objective is to methodically gather network performance log data using G-NetTrack Pro, an advanced tool that records comprehensive network parameters required for later analysis. This data undergoes extensive preprocessing to guarantee its quality and reliability. This involves removing outliers to prevent biased results, filling in missing values to preserve the integrity of the dataset, improving the data to enhance the accuracy of models, and adjusting the features to ensure that all variables are appropriately weighted during the model's learning process.

Expanding upon this basis, the project advances towards constructing prognostic models, each of which employ one the three selected machine learning algorithms: K-nearest neighbor (KNN) clustering, logistic regression, and random forest are three machine learning algorithms. Each algorithm is selected based on its distinct strengths and how it generates diverse perspectives on data, thereby enabling a more comprehensive analysis of predictive capabilities. Models undergo rigorous evaluation based on key performance metrics to determine their accuracy and reliability. This evaluation process helps identify the most effective systems for predicting network performance.

Next, network time series forecasting models, more specifically the ARIMA (AutoRegressive Integrated Moving Average) and FBProphet models, are utilized to help aid in network throughput predictions. Both are well-known for their proficiency in predicting temporal patterns. These fashion trends are utilized in community performance analysis to accurately forecast future community conditions. The performance of these forecasting models is subsequently evaluated primarily based on their Root Mean Square Error (RMSE) values, which provide a quantitative measure of their accuracy. The final phase of the project involves a comprehensive comparative assessment of all sophisticated models. This comparative assessment specifically examines the predictive precision and dependability of both the machine learning and time series forecasting models, providing a nuanced understanding of the strengths and limitations of each model. The assessment examines key performance metrics to identify models that perform exceptionally well and provides insights on incorporating various modelling techniques to enhance overall network performance optimization.

Through a systematic approach, the project aims to address crucial aspects to provide practical insights and effective solutions that significantly enhance the performance of 4G networks. The comprehensive analysis provides valuable insights that will greatly benefit community operators, policymakers, and researchers. It leverages the capabilities of modern telecommunications infrastructure and ensures that user demands for high-speed data and consistent quality of service are met with innovative and effective solutions. This challenge specifically focuses on improving the performance of 4G networks by using a comprehensive approach that includes collecting data, preprocessing it, creating predictive models, and conducting comparative assessments.

The utilization of advanced machine learning and time series forecasting techniques allows for robust solutions and actionable insights that guarantee significant enhancements in network performance and customer connectivity. The comprehensive evaluation and results are positioned to benefit network operators, policymakers, and researchers, thus enhancing the capabilities of current telecommunications infrastructure to meet the growing demand for high-speed data and reliable QoS. The diagrammatic flow-chart of the overall prediction process of the project is shown below in figure 2.1.

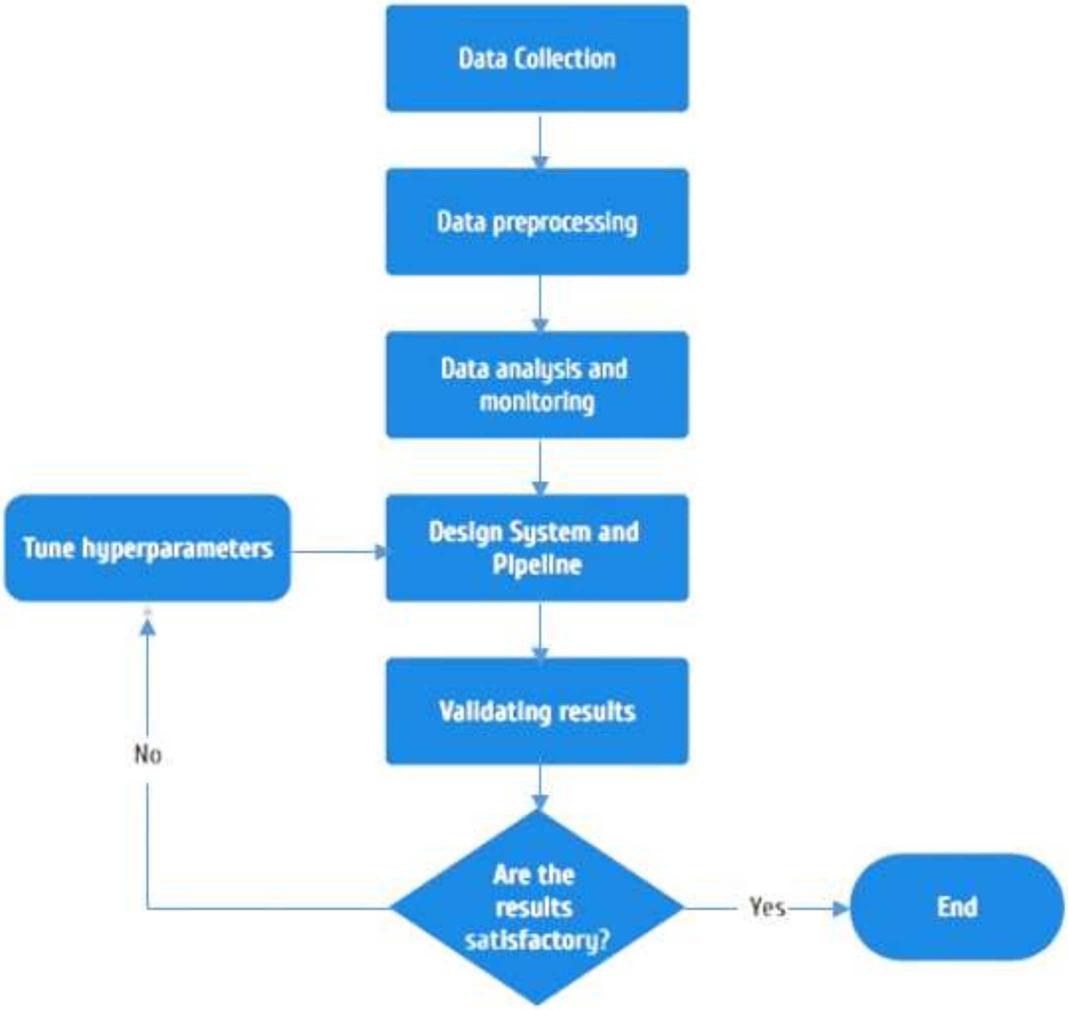

Figure 2.1: Downlink throughput prediction process

Chapter 3: Model Formulation

3.1. Data Acquisition

The dataset utilized for this project was carefully gathered using G-NetTrack Pro, a powerful network data logging software known for its extensive capability to capture metrics related to network performance. To guarantee a dataset that encompasses a wide range of characteristics and accurately reflects the population, numerous expeditions were conducted in Dhaka, Bangladesh, visiting various locations. Every outing lasted at least one hour and was carried out systematically. The aforementioned areas encompass Gazipur, Uttara, Banani, Baridhara, Dhanmondi, Mirpur, Mohammadpur, Rampura, and Bashundhara. The choice of these locations was intentional, including both densely populated residential areas and heavily trafficked transportation routes. The implementation of this strategic planning ensured that the collected data encompassed a broad range of network performance scenarios, including highly populated urban centers and sparsely populated suburban areas.

Dhaka, the capital city of Bangladesh, is renowned for its exceptionally high population density, surpassing 21 million individuals in the metropolitan area, making it one of the most densely populated cities in the world. The high population density presents considerable obstacles for telecommunications networks because of the extensive data congestion caused by the sheer volume of users. The urban environment's lively and energetic atmosphere provides a rich and abundant dataset for analysis, resulting in valuable insights into network performance under different circumstances. The dataset in Dhaka contains a significant number of data points that represent various usage patterns, signal strengths, and network loads. This outcome is a direct consequence of the substantial user base and the extensive geographic coverage in the region. The presence of a wide range of different elements is essential for the creation of strong and precise predictive models. The selected regions consist of a mixture of various socioeconomic zones and types of infrastructure. Gazipur and Uttara are experiencing rapid development, marked by the creation of thriving commercial and residential areas. Nevertheless, Banani and Baridhara are renowned for their affluent residential communities and bustling commercial districts. Dhanmondi and Mirpur are vibrant areas known for their mix of residential, commercial, and institutional developments, including educational institutions like schools and colleges, as well as medical facilities. The diverse array of activities leads to different patterns of data traffic throughout the day. Mohammadpur and Rampura are well-known for their densely populated residential areas and lively local markets, which contribute to the diversity of the dataset and include common urban challenges. Bashundhara is a highly regarded residential and commercial area that offers valuable information on network performance in carefully designed urban settings.

The excursions were conducted at regular intervals throughout the day, from 9AM to 12PM, then from 2PM to 4PM, and later in the evening from 5PM to 8PM. The temporal distribution is crucial for capturing the variations in network performance throughout the day, including peak usage hours when users are most active, as well as periods of reduced activity. The collected data demonstrates the relationship between network performance and user density and movement patterns, as observed by tracking transportation routes and residential areas in this densely populated city. Greater population density results in heightened data consumption, particularly in residential areas where users are more prone to streaming videos, engaging in online gaming, and participating in video conferencing. These activities require a substantial amount of bandwidth and a reliable internet connection.

The inclusion of a wide range of socioeconomic and infrastructural factors from Dhaka in the dataset improves the reliability and accuracy of the predictive models created in this project. The distinct attributes of each location contribute to a comprehensive comprehension of network behavior, facilitating the creation of models that can be efficiently utilized in various scenarios. The systematic methodology employed for data collection serves two purposes: first, it aids in the creation of precise predictive models; second, it guarantees that the outcomes and solutions derived from this project are relevant and practical in real-life situations. Ultimately, this approach enhances network performance and user experience on 4G networks in Dhaka.

Overall, the careful and systematic collection of data using G-NetTrack Pro in different locations in Dhaka and Gazipur has provided a strong and comprehensive dataset for this project. The strategic planning includes covering both high-traffic urban facilities and residential areas, as well as capturing data at different times of the day, ensuring a comprehensive and diverse dataset. The variability of information factors is essential for the development of accurate and dependable predictive models for network performance. The comprehensive understanding of community behavior in diverse socioeconomic and infrastructural settings enhances the relevance and efficiency of predictive models. The primary objective of this mission is to significantly improve the performance and user experience of the 4G network in Dhaka. This will address the challenges caused by the high population density and diverse usage patterns in the area.

3.1.1. Parameters of the Dataset:

The dataset comprises multiple network parameters essential for assessing and predicting network performance. By conducting a simple correlation analysis of downlink throughput with the rest of the dataset, it was found that the following parameters had the highest match with the throughput, and as a result were selected as the feature variables used to make downlink throughput predictions:

- *RSRP (Reference Signal Received Power):* Specifies the magnitude of the received signal's power. RSRP, which stands for Reference Signal Received Power, is a crucial parameter in LTE networks. It is utilized to evaluate the strength and reliability of the connection, as well as to make decisions regarding cell selection and handover. It aids in comprehending the extent of signal coverage and identifying areas where the signal strength may be insufficient, thus requiring optimization.
- *SNR (Signal-to-Noise Ratio):* Signal quality is assessed by comparing the amplitude of the desired signal with the amplitude of the background noise. A higher signal-to-noise ratio (SNR) signifies a signal that is more pristine and robust, which is essential for sustaining elevated data transmission speeds and diminishing the occurrence of errors. The signal-to-noise ratio (SNR) is of utmost importance in urban areas due to the potential degradation of network performance caused by interference from other signals.
- *RSSI (Received Signal Strength Indicator):* It quantifies the magnitude of the signal received, by summing the power of all received signals and noise. RSSI is utilized in diverse wireless communication protocols and offers a comprehensive assessment of the overall signal environment, encompassing the desired signal, interference, and background noise.
- *Downlink and uplink throughput:* This data represents the rates at which information is transferred when downloading and uploading. Downlink throughput refers to the rate at which data is received by the user device from

the network, while uplink throughput refers to the rate at which data is sent from the user device to the network. These metrics are crucial for assessing the user experience, particularly for applications that require a large amount of bandwidth, such as video streaming, online gaming, and video conferencing.

- *Geographical information:* The latitude and longitude coordinates indicate the precise locations where the data was gathered. Geographical tagging enables the analysis of network performance about location, facilitating the identification of regions with inadequate coverage or significant interference. It allows for the generation of coverage maps and simplifies focused optimization efforts in particular areas.
- *Timestamps:* The exact timestamps when the data was recorded, spanning a wide timeframe from 9 AM to 1 AM. By including timestamps, it becomes possible to conduct a temporal analysis of network performance, which involves capturing and studying the variations that occur throughout the day. The temporal dimension is crucial for comprehending periods of peak usage, which can exert pressure on network resources, and for identifying times of low demand when network performance may be enhanced.

To obtain a complete understanding of the network's performance throughout the day, it is necessary to implement a data collection strategy that considers both peak and off-peak hours and covers all aspects. The dataset accurately captures the dynamic nature of urban telecommunications by incorporating information from diverse usage patterns and network conditions. This is made possible by the collection of information at various intervals throughout the day.

3.1.2. Relevance to Urban Telecom:

Dhaka has a dense population, necessitating a network capable of efficiently handling a substantial volume of connections simultaneously. Although this scenario poses several challenges, it does offer valuable insights into the network's performance under heavy load conditions. To achieve optimal capacity and ensure reliable service, this information is essential.

The inclusion of residential areas and transportation routes ensures that the dataset can capture a diverse range of usage scenarios. One example of this principle being implemented is the incorporation of residential areas. Residential areas may encounter elevated data consumption in the evening when individuals are present at their homes. However, transportation routes may encounter varying demands because of commuters traversing different network cells than their usual ones. Infrastructure development in Dhaka varies across different areas, with some areas having densely populated high-rise buildings while others have more open spaces. This has had a significant influence on the city's infrastructure. Collecting data across diverse environments is crucial for developing robust models that effectively account for the impact of signal propagation and interference patterns. This is due to the impact of this diversity. The dataset provides an indirect depiction of how environmental factors, such as weather conditions, construction materials, and urban topography, impact signal quality and network performance. This is due to the inclusion of data about these factors in the dataset. For example, tall buildings can cause signal reflections and shadowing, which can impact parameters like RSRP and SNR.

3.1.3. Applications of the Dataset:

The creation of an accurate dataset is one of the focal points of the project. Creating a dataset that is complete and reliable ensures that all models based on the dataset work as effectively as possible. The finalized dataset has the following applications in the project:

- *Network optimization:* The collected data can be used to optimize network parameters, improving coverage and capacity. By analyzing the geographical and temporal variations in network performance, operators can deploy resources more efficiently and enhance the user experience.
- *Predictive modelling:* With the rich set of parameters, machine learning models can be trained to predict network performance metrics such as downlink throughput. These predictions can help in proactive network management, allowing operators to anticipate and mitigate potential issues before they impact users.
- *Service quality enhancement:* Understanding the factors that influence key metrics like SNR and throughput can lead to better service quality. For example, if data shows that certain areas consistently experience poor performance, targeted interventions such as installing additional cell sites or adjusting antenna configurations can be implemented.
- *Policy and planning:* Insights from the dataset can inform policy decisions and urban planning efforts. For example, areas with consistently poor network

performance might be prioritized for infrastructure investments, and planning new developments can consider the impact on telecommunications networks.

To summarize, the dataset collected for this assignment using G-NetTrack Pro is comprehensive and contains a wide range of network parameters that are crucial for detailed analysis and predictive modelling. Spanning a diverse range of locations in Dhaka and Gazipur, and capturing data at various times throughout the day, this dataset provides a strong foundation for analyzing and enhancing the overall performance of the 4G network. The incorporation of essential metrics such as RSRP, SNR, RSSI, and throughput, along with accurate geographical and temporal statistics, enables precise spatial and temporal analysis. The dense population and diverse infrastructure of Dhaka present unique challenges and opportunities for telecommunications networks. The gathered data showcases the ever-changing urban environment, capturing a wide range of usage scenarios and community situations. The presence of this variety is crucial to developing precise predictive models and highly effective optimization techniques. By utilizing those observations, network operators can improve their capacity, and range, and deliver superior quality, ensuring dependable and high-speed connectivity for users.

Furthermore, the insights derived from this dataset can provide valuable information for making decisions about coverage and strategic planning, aiding in the prioritization of infrastructure investments and improving overall community performance. The projects' thorough methodology for gathering and analyzing data not only aids in the development of predictive models but also ensures that the resulting solutions apply to real-world situations. This ultimately leads to improved network performance and customer satisfaction in the densely populated regions of Dhaka and Gazipur.

3.2. Data Preprocessing

Data preprocessing is a crucial and fundamental step in preparing the dataset for analysis and subsequent model development. This step ensures that the information is seamless, consistent, and properly formatted, thereby making it suitable for precise and dependable predictive modelling [14]. In the absence of comprehensive preprocessing, the unprocessed data can result in misleading conclusions and subpar performance of the model due to the presence of noise, inconsistencies, and missing values. The main steps involved in data preprocessing for this mission are outlier removal, missing value imputation, data augmentation, and feature scaling. Each step serves a crucial role in transforming raw data into a format that can be effectively utilized to enhance robust predictive models. Hence, data preprocessing is an essential stage in preparing the dataset for analysis and model enhancement. This procedure guarantees that the information is free from errors and consistent. The main steps involved in data preprocessing for this project include removing outliers, imputing missing values, augmenting data and scaling features. Each of those steps serves a vital role in transforming unprocessed data into a format that can be effectively utilized to develop predictive models.

3.2.1. Different preprocessing techniques applied:

The careful implementation of preprocessing techniques optimizes the dataset for building strong prediction models. By incorporating these methods, it is possible to make sure that the models built with this data will provide trustworthy insights and forecasts, improving the 4G network's performance and the user experience in Dhaka and Gazipur regions. Extensive statistical preprocessing improves the accuracy and stability of models, providing a solid foundation for further analysis and optimization methods.

- *Outlier detection and removal:* The first preprocessing technique applied was outlier detection and its removal from the obtained dataset. Outlier detection is the process of finding out data points that vary significantly from the rest. The existence of outliers in a dataset for predictive modelling will heavily affect the ability of the model to make accurate predictions, and therefore it is essential to ensure that outliers in the dataset are removed before it is passed on to the model. For detection of outliers, the z-score method was used. The z-score indicates where a data point is from their mean value by measuring the multiples of standard deviation above or below the mean. The z-score of a data point can be calculated using the following equation:

$$z = \frac{x - \mu}{\sigma}$$

  where x is the data point, μ is the mean of all data points and σ is the standard deviation.

- *Missing value imputation:* Imputation of missing values from the dataset was the second preprocessing method employed. When trying to create predictive models, missing values in a dataset are a common cause of inaccuracy. Most machine learning algorithms simply cannot function with missing values in a dataset and will either stop the run or produce incorrect results. Imputation was used to prevent these problems from happening. The mean value of the remaining feature data points was used to fill in the missing values in each feature column.
- *Feature scaling:* Feature scaling was the next preprocessing technique that was applied. This technique is generally required when the dataset has features that have different ranges, which is especially important in models that rely on the distances between features. This is due to the features with bigger ranges will often have a greater influence on model outcomes than features with lower ranges. Therefore, the preprocessing technique of standardization is applied to the dataset. Standardization is a scaling technique that transforms the distribution of a feature so that it has zero mean and a standard deviation equal to one. In the case of the project, the StandardScaler function of the sklearn library was used [15]. This function is based on the theory that the dataset variables, whose values lie in differing ranges, do not have equal contribution to the models' fit parameters and training function, and may lead to a bias in the predictions made by the model. By eliminating the mean from the features and scaling them to unit variance, features are standardized using this function.
- *Dataset formatting:* The dataset is later finalized by setting the correct format to every feature variable involved. This includes setting the variables in order so that the models make the most effective use of the dataset. Setting the correct format of the dataset also enables the use of geographical visualization tools such as Google Earth to be used, allowing for a visual understanding of how different parameters change and their subsequent effect on the downlink throughput.

3.2.2. Benefits of data preprocessing:

Data preprocessing gives several advantages, essential for growing powerful and dependable predictive fashions. These benefits are multifaceted, improving the excellence of the dataset and the performance of the fashions. Here are some key blessings, extended with more applicable facts:

- *Improved model performance:* Clean and nicely prepared facts significantly enhance the version's overall performance. By disposing of noise and inconsistencies, preprocessing guarantees that the models are educated on correct and relevant information, main to greater precise and reliable predictions. Models built on preprocessed statistics typically exhibit higher accuracy and higher generalizability when applied to new information.
- *Enhanced data quality:* Preprocessing improves the overall great of the dataset by addressing mistakes, duplicates, and inconsistencies. This enhancement ensures that the facts as they should reflect the actual global conditions its objectives to version, presenting a strong foundation for evaluation. High-exceptional statistics are essential for deriving meaningful insights and making informed choices.
- *Better handling of missing values:* Missing statistics is a common trouble that could introduce bias and reduce the reliability of the models. Proper imputation techniques, together with imply, median, or mode imputation, make sure that missing values are handled without distorting the dataset. This cautious control of lacking values keeps the integrity of the statistics and supports strong model schooling.
- *Robustness to outliers:* Unusual data points, known as outliers, can have a significant impact on statistical analyses and predictions made by models if they are not appropriately handled. Preprocessing ensures that the results are not negatively affected by outliers by identifying and managing them, either by removing or adjusting them. Ensuring the accuracy and stability of models is crucial, particularly in datasets with varied conditions.
- *Scalability and consistency:* Feature scaling, which encompasses both normalization and standardization, guarantees that all features have an equal impact on the model's learning procedure. This step is crucial for algorithms that are highly responsive to the magnitude of input data, such as K-Nearest Neighbors and neural networks. Preprocessing facilitates consistent performance across various data ranges and improves the model's scalability with larger datasets by standardizing features to a common scale.

- *Reduced complexity:* By cleaning and organizing the data, preprocessing reduces the complexity that models need to handle, making them more efficient and easier to train.
- *Enhanced interpretability:* Well-preprocessed data can make the results of the analysis more interpretable, allowing for clearer insights and more actionable recommendations.
- *Time and resource efficiency:* Proper preprocessing can save significant time and computational resources in the later stages of model training and evaluation, as it minimizes the need for extensive data cleaning and manipulation during model development.

Data preprocessing is a crucial stage in the creation of predictive models. It transforms unprocessed data into a format that is easy to access, strong, and can be searched, solving issues such as unusual data points, incomplete data, and measurements of different characteristics. This step is important for ensuring dataset satisfaction and enhancing reliability, significantly enhancing the accuracy and resilience of prediction models. It effectively represents conditions and guarantees the absence of noise and inconsistencies. Effective handling of records through preprocessing not only enhances overall operational efficiency but also guarantees that insights derived from information are significant and can be acted upon. Preliminary data generation is an essential step in fully harnessing the potential of predictive modeling. It enables the development of new, realistic, and reliable predictions that can effectively condense data for informed decision-making and optimized network performance. Figure 3 illustrates a visualization of the downlink throughput around Dhanmondi on Google Earth.

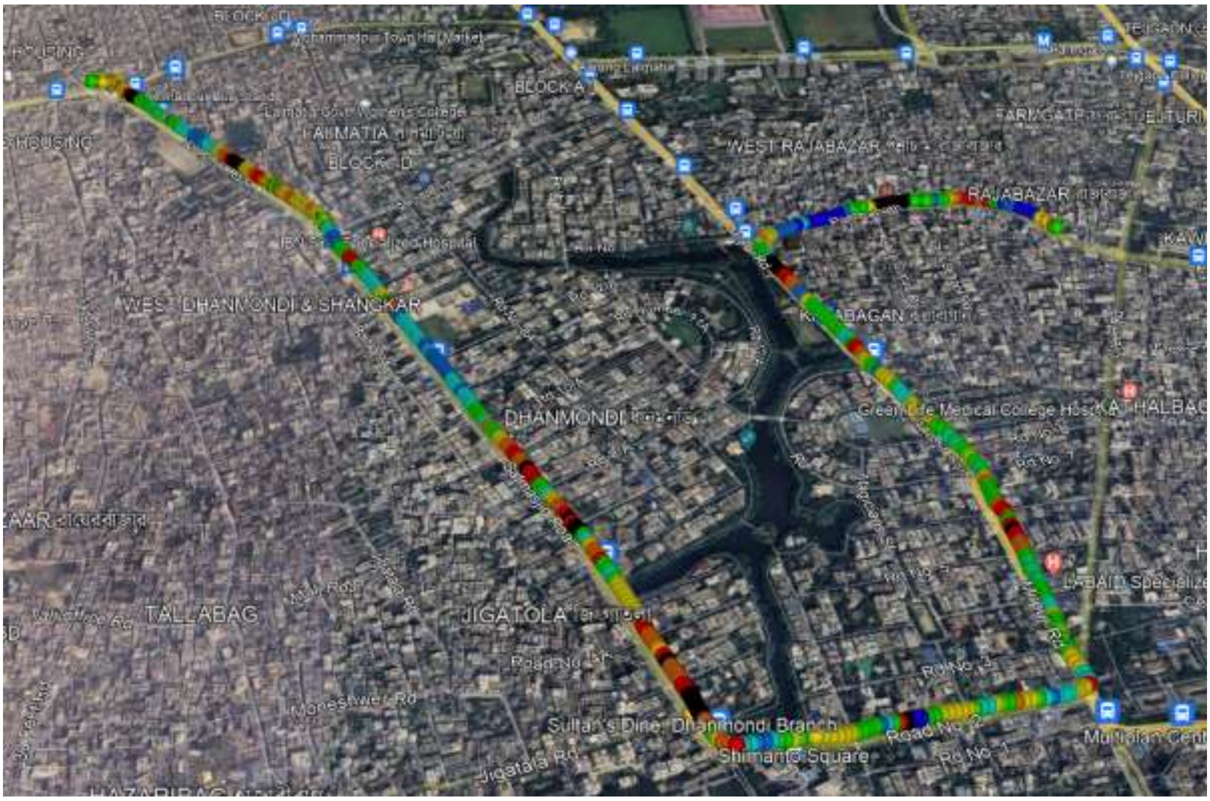

Figure 3.1: Downlink throughput collected around Dhanmondi

3.3. Development of Throughput Predictive Models

The primary focus of this mission is to enhance and assess the system learning prediction models to optimize the performance of 4G networks and improve the Quality of Service (QoS) in densely populated areas of Gazipur and Dhaka. The main objective is to develop predictive models that can accurately forecast the average performance metrics of a network, enabling proactive measures to uphold and enhance carrier excellence. This section covers the selection, implementation, and assessment of three machine reading algorithms: K-nearest neighbors (KNN), Logistic Regression, and Random Forest. Each of those algorithms was applied to a preprocessed dataset that included multiple community parameters, such as RSRP, SNR, RSSI, and throughput rates. The consequences of these models have been compared to determine the optimal set of rules for predicting network average performance and addressing the specific challenging conditions presented by high data traffic and diverse user environments in Dhaka.

3.3.1. K-nearest Neighbors:

K-Nearest Neighbors (KNN) is a fundamental and highly adaptable machine learning algorithm widely employed for both classification and regression tasks. The fundamental principle of KNN is based on the concept of proximity or similarity, where the classification or prediction of a statistical point is determined by the labels or values of its closest neighbors in the feature space. The proximity-based approach enables KNN to efficiently detect patterns and relationships in the data by examining the nearest data points to the target sample. The strength of KNN lies in its simplicity and efficacy, making it an extremely intuitive algorithm suitable for a wide range of applications. The system operates on a reliable mechanism where it stores all available time data and generates predictions based on the 'k' nearest data points. Although KNN is computationally demanding, especially when dealing with large datasets, and is sensitive to characteristic scaling, it is still highly valued as a tool in the machine learning toolkit. The reason for this is the strong and reliable performance, the simplicity of implementation, and the ability to effectively handle both classification and regression problems. The effectiveness of KNN is based on its ability to utilize the closeness of similar data points to produce precise predictions and classifications. By prioritizing the closest connections, KNN can provide valuable insights and solutions that are particularly applicable to the specific statistical context.

In summary, the KNN algorithm is essential in the field of machine learning due to its intuitive nature, flexibility, and strong overall performance. The ability of this method to classify and predict based on the proximity of data points ensures its reliability and effectiveness for various analytical and predictive tasks. In the context of this project, KNN can be applied to predict downlink throughput in 4G networks based on various network parameters. Here's how KNN is utilized:

1) *Data preparation:* Before applying KNN, the dataset is preprocessed to ensure it is clean and standardized. This involves steps like outlier removal, missing value imputation, and feature scaling.
2) *Choosing 'k':* The optimal value of 'k' is determined using cross-validation. This helps in balancing the bias-variance tradeoff and ensures the model generalizes well on unseen data.

3) *Distance metric:* Euclidean distance is typically used to measure the similarity between data points. Given the scaled features, this metric calculates the straight-line distance between points in the multi-dimensional space.
4) *Model training:* In KNN, the training phase is straightforward as the model essentially stores the training dataset. The computational load comes during the prediction phase, where the algorithm calculates distances to determine the nearest neighbors.
5) *Prediction:* For a new data point, the algorithm identifies the 'k' nearest neighbors from the training set and predicts the downlink throughput based on these neighbors. In a regression context, this might involve taking the average throughput values of the nearest neighbors.
6) *Evaluation:* The performance of the KNN model is evaluated using metrics such as Root Mean Squared Error (RMSE) to assess how well the predicted values match the actual values.

For example, if we are predicting the downlink throughput for a specific time and location, KNN will:

- Identify the most similar 'k' observations from the data set based on parameters such as RSRP, SNR, RSSI, and geographic information.
- Calculate the downlink throughput of these ‘k’ neighbors.
- Use this average as the predicted downlink throughput at a given time and location.

To summarize, the straightforwardness and efficiency of KNN make it an excellent option for forecasting downlink throughput in 4G networks. KNN utilizes the proximity of similar data points to make dependable predictions, thereby enhancing the efficiency of the network.

3.3.2. Logistic Regression:

Logistic Regression is a robust and widely used algorithm, particularly suited for binary classification tasks. This algorithm operates by modeling the probability that a given data point belongs to one of two classes. Logistic regression predicts a probability value that falls between 0 and 1. This probability is then used to classify the data point into one of the two classes.

The logistic regression model is based on the logistic function, also known as the sigmoid function, which maps any real-valued number into a value between 0 and 1. The sigmoid function is defined as:

$$\sigma(x) = \frac{1}{(1 + e - x)}$$

where x is a linear combination of the input features. The model estimates the coefficients of this linear combination by maximizing the likelihood of the observed data.

Logistic regression is highly effective in real-world applications, primarily because of its straightforwardness and effectiveness. The model assumes a linear correlation between the input features and the logarithm of the odds of the outcome, which simplifies the implementation and interpretation process. Moreover, it offers valuable observations regarding the impact of each characteristic on the likelihood of the result, as indicated by the estimated coefficients. It is a fundamental and powerful machine learning algorithm that is commonly used for binary classification problems. It offers a straightforward approach to solving these types of problems. The applications of this technology are wide-ranging, encompassing various fields such as medical diagnosis, credit scoring, and marketing. The primary objective is to forecast the probability of a specific event or outcome by analyzing input features.

Logistic regression excels in its capacity to effectively handle diverse input features, encompassing both continuous and categorical variables. Moreover, it exhibits a lower susceptibility to overfitting in comparison to more intricate models, particularly when employing regularization techniques like L1 (Lasso) or L2 (Ridge) regularization.

3.3.3. Random Forest:

Random Forest is an ensemble learning method that is employed for classification, regression, and other tasks. It operates by constructing a few selected trees at the beginning of the learning process and generating the mode of the classes (classification) or the mean prediction (regression) of the character trees. It is an extension of the choice tree set of rules and targets that is designed to address the tendency of selection trees to overfit the education set.

Random Forest can generate numerous decision trees during the training process, which are subsequently combined to generate a final prediction. Bootstrap aggregating, which is occasionally referred to as "bagging," is a method of generating multiple decision trees from distinct regions of the dataset. The ensemble method operates in this manner. A subset is randomly selected from the original dataset and substituted. This ensures that each tree is trained on a slightly distinct set of data. Additionally, each tree is equipped with a distinct set of attributes, which further distinguishes them from one another. This type is crucial because it reduces the link between trees, which in turn enhances the model's functionality.

The algorithm divides data points into branches by employing conditions that are based on feature values. Therefore, a tree-like structure is established to illustrate the potential outcomes and the choices that can be made. Random Forest examines the predictions from all the decision trees and selects the group that receives the most votes to organize items into groups. This is accomplished by utilizing the "wisdom of the crowd" to make accurate predictions. The algorithm determines the optimal solution by averaging the predictions generated by the trees during regression tasks. This prevents the results from being distorted by values that are either too high or too low from any one tree. By averaging their predictions, this ensemble method mitigates the risk of overfitting. This is a common occurrence with single decision trees. Random Forest can be employed for a diverse array of tasks when presented with a large dataset or a space with numerous dimensions. Even when there are significantly more features than observations, it continues to function effectively. Additionally, the algorithm can prioritize the significance of each feature in predicting the target variable. This informs you of the significance of the features. Additionally, this assists in determining which features to employ and how the data is interconnected. Random Forest is not only consistent and accurate, but it is also relatively straightforward to configure and operate. To achieve optimal speeds, it is possible to modify critical parameters, such as the number of trees in the forest and the number of features to examine at each split. The algorithm is capable of coping with missing data and continuing to operate in the absence of values. This renders it even more advantageous in practical use.

Random Forest is an excellent algorithm that is user-friendly and can be customized to meet the requirements of various applications. This being the case, it is applicable for both regression and classification. It can make accurate predictions in a wide range of fields and can easily adjust to new data due to its ability to work with groups of data.

Random Forest is implemented in this project to forecast downlink throughput in 4G networks by utilizing a variety of network parameters. The data is initially preprocessed to address outliers, impute missing values, augment data, and scale features, thereby guaranteeing a standardized and clean dataset for model training.

Different bootstrapped subsets of the data are employed by Random Forest to construct multiple decision trees during the training process. The model can capture diverse patterns and variability in the network parameters by utilizing a random subset of features in each tree. RSRP, SNR, RSSI, geographical information, and timestamps are among the most significant features that influence downlink throughput, as identified by this process. The model generates predictions for new data points by averaging the predictions from each tree after it has been trained. To evaluate the accuracy and reliability of the predictions, the Random Forest model is assessed using metrics such as Root Mean Squared Error (RMSE). The model's predictions provide valuable insights that aid in the identification of potential bottlenecks and areas for optimization in the network, thereby improving user experiences and enhancing QoS in the 4G network.

The objective of this project is to create a predictive model for downlink throughput that is both accurate and robust, thereby effectively addressing the complexities and variability that are inherent in 4G network data. This will be achieved by utilizing Random Forest. The model's reliability is guaranteed by the ensemble approach, which captures a diverse array of patterns and interactions in the data.

3.4. Development of Time Series Forecasting Models

Time series forecasting is a crucial element of predictive modeling. This is especially true when dealing with dynamic systems, such as telephone networks, that experience changes over time. This project employs forecasting methodologies to anticipate future fluctuations and patterns in network performance in the highly populated cities of Gazipur and Dhaka. The main goal is to construct models that can precisely forecast the future behavior of networks. Therefore, proactive strategies can be employed to improve the Quality of Service of 4G networks. FBProphet and ARIMA are two well-known approaches for utilizing time series data to make predictions. Each method is highly advantageous for this arduous task as it excels in carrying out distinct functions.

Facebook created a flexible predictive tool called FBProphet. The system has the ability to analyze time series data that displays prominent seasonal patterns and multiple cycles, such as daily, weekly, and yearly cycles. The most advantageous time to employ it for making predictions is when the data demonstrates a regular pattern with occasional sudden fluctuations. FBProphet is specifically designed to be intuitive and approachable, enabling individuals without a background in statistics to effortlessly utilize and understand it. Time series data can be classified into three main categories: trend, seasonality, and holidays. FBProphet can effectively detect patterns in data and produce accurate predictions, even when there are missing data points and outliers, by modeling these elements separately.

AutoRegressive Integrated Moving Average (ARIMA) is a well-established statistical method that has been extensively used for a significant period of time to predict time series data [16]. The ARIMA model is highly proficient in identifying different elements of time series data, including trends and autocorrelations. The composition comprises three fundamental components: autoregression (AR), differencing (I), and moving average (MA). Differentiation is a process that removes patterns to make the time series stationary. Autoregression is a method that uses historical data to forecast future values. The moving average filters out extraneous fluctuations in the data. ARIMA is beneficial because it is dependable and capable of accurately predicting future outcomes by effectively representing the autocorrelation pattern of the data.

The prediction methods used in this project are applied to estimate the downlink throughput. Downlink throughput is an essential measure for evaluating the efficiency of a network. The provided data comprises a chronologically ordered set of measurements of the rate at which data is transmitted from a central source to multiple locations in Dhaka and Gazipur. The data exhibits a diverse array of network conditions and user behaviors. Network operators can effectively forecast and resolve potential issues by accurately predicting future data transfer rates. This guarantees the highest level of performance. The choice was made to employ FBProphet and ARIMA because of their mutually beneficial capabilities. FBProphet can efficiently handle both seasonal patterns and sudden changes, making it well-suited for analyzing intricate trends typically seen in network performance data that span across multiple time periods. ARIMA guarantees the precision of short-term forecasts by proficiently modeling dynamic structures. Timely network management is essential. These models work together to create a comprehensive framework for predicting future outcomes, incorporating the most efficient elements from each model.

When faced with the complexity of time series data, it is essential to utilize a range of forecasting methodologies and see what best fits the actual scenario. The incorporation of both FBProphet and ARIMA in this project highlights its importance. The project aims to integrate the most efficient elements from both models to offer accurate and valuable insights into the operation of future networks. Implementing this measure in densely populated urban areas in Dhaka and Gazipur will help optimize the performance of 4G networks, allowing them to support a greater number of users.

3.4.1. FBProphet:

FBProphet is a forecasting tool developed by Facebook's Data Science group. It is designed to be simple, flexible, and easy to understand for time series forecasting. The tool breaks down time collection statistics into three primary components: trend, seasonality, and holidays. The fashion is modeled using a piecewise linear or logistic growth curve, while seasonality is accounted for using a Fourier series. Person-defined lists of crucial dates are used to address the consequences of holidays. FBProphet automatically adapts the additives in the data, considering outliers and missing values, to generate reliable forecasts with confident intervals. The project applies FBProphet to analyze the downlink throughput in 4G networks across diverse areas in Dhaka and Gazipur. FBProphet effectively captures the daily, weekly, and annual patterns in network performance data by utilizing its ability to handle multiple seasonality and sudden changes. The advantages of utilizing FBProphet in this endeavor include its automated trend and pattern detection, as well as the generation of interpretable outcomes with uncertainty intervals. These features assist in accurately assessing the confidence of predictions and making informed decisions for network optimization. FBProphet, commonly known as Prophet, is an open-source forecasting tool developed by Facebook's Data Science group. This tool is specifically designed to handle time series statistics that exhibit strong seasonal patterns and have multiple seasons of historical data. The Prophet software is particularly suitable for corporate forecasting tasks, such as predicting future sales, inventory levels, and website traffic. FBProphet operates by decomposing time series data into three main components: trend, seasonality, and holidays.

1) *Trend:* Prophet models the trend component using either a piecewise linear model or a logistic growth model. The piecewise linear model allows for changes in the rate of growth, while the logistic growth model is useful when there is a saturating growth trend.
2) *Seasonality:* This component captures the repeating patterns in the data that occur at regular intervals, such as daily, weekly, or yearly cycles. Prophet uses Fourier series to model seasonal effects, which allows it to capture complex seasonal patterns.
3) *Holidays:* Prophet can incorporate the effects of holidays and special events by allowing users to provide a list of dates that are important for their forecasting. This is particularly useful for businesses that experience significant variations around holidays.

The Prophet algorithm applies these components to historical data, considering any anomalies or gaps in the data. The additive model incorporates the trend, seasonality, and holiday effects. The algorithm is specifically designed to be resilient to data gaps and changes in the pattern, rendering it extremely dependable for practical scenarios where data can be untidy.

This project utilizes FBProphet to forecast downlink throughput in 4G networks across various regions in Dhaka and Gazipur. The objective is to identify and quantify the variations in community performance, which is crucial for optimizing network operations and ensuring a high standard of service quality. The process involved the following steps:

- *Data preparation:* The network performance data, including downlink throughput, was collected and preprocessed. This involved handling missing values, removing outliers, and ensuring the data was in a time series format suitable for Prophet.
- *Modeling trends and seasonality:* Prophet was used to model the long-term trends and seasonal patterns in the data. Given the nature of telecommunications usage, daily and weekly seasonality were particularly important.
- *Forecasting and evaluation:* Prophet was then used to generate forecasts for future downlink throughput. The performance of these forecasts was evaluated using metrics such as RMSE (Root Mean Squared Error) to ensure accuracy and reliability.

FBProphet offers a multitude of advantages that make it especially well-suited for analyzing and predicting the performance of 4G networks. One of its primary advantages is its capacity to consistently identify and represent underlying trends and seasonal patterns within the data. This feature significantly reduces the need for extensive manual adjustments, making it easier to begin forecasting. The ability of Prophet to handle missing data and effectively handle outliers is crucial, especially considering that real-world datasets often contain such imperfections. Moreover, the flexibility of Prophet in integrating vacations and special events allows for more precise modeling of periods during which network usage may deviate from normal patterns, thereby enhancing the reliability of the forecasts.

Another significant advantage of Prophet is its capacity for providing interpretable outcomes. It provides forecasts alongside periods of uncertainty, which simplifies the interpretation of predictions and understanding of the associated confidence level. This feature is primarily advantageous for making well-informed decisions based on the forecasts. Moreover, the device's user-friendliness is remarkable. Prophet is designed to be user-friendly, with a reliable API that can be easily utilized by both experienced data scientists and individuals with limited technical knowledge. This enhanced accessibility expands the tool's suitability for a wide range of users.

This utilization of FBProphet aims to help expand reliable and interpretable models for forecasting 4G network performance. These fashion trends will ultimately contribute to more efficient community management and optimization strategies. Precisely predicting network performance is essential for strategic planning and efficient allocation of resources, guaranteeing that the network can effectively manage fluctuating workloads and uphold a superior level of service for users. FBProphet's extensive benefits make it a valuable tool for improving the performance and reliability of 4G networks.

3.4.2: Auto-Regressive Integrated Moving Average (ARIMA):

The acronym ARIMA is an abbreviation that stands for "Auto-Regressive Integrated Moving Average," which is a model that is widely recognized and frequently used for the purpose of forecasting time series data. It is highly regarded for its ability to accurately capture a variety of characteristics of a dataset, including noise, patterns, and trends, among other data characteristics. There are many other data characteristics that it can accurately capture. There are three components that make up the ARIMA model. The first component is the autoregressive component (p), which is responsible for capturing the relationship between an observation and previous observations. The second component is the integrated component (I), which is responsible for representing the difference of raw observations to achieve stationarity in the time series. The third component is the moving average component (q), which is responsible for modelling the relationship between an observation and the residual error that is derived from a moving average model using previous observations. The autocorrelation function is typically utilized to perform the estimation of these components, and the maximum likelihood estimation (MLE) method is utilized in order to ascertain the values of these individual components. The stationarity of the data is a requirement for the ARIMA model, which means that its statistical properties must remain unchanged repeatedly. This is a prerequisite for the model. The ARIMA method is an extremely effective technique for making short-term forecasts of datasets that exhibit seasonality and trends. Even though it has a few significant drawbacks, the ARIMA method is an extremely efficient technique. As a result of this, it is a dependable choice for dealing with a wide range of applications that involve time series.

The use of the ARIMA model helps to make projections regarding the downlink throughput in 4G networks. The dataset is subjected to stringent preprocessing to guarantee stationarity. This guaranteeing stationarity makes it easier for ARIMA to accurately model the linear dependencies that are present in the data. To acquire an understanding of the patterns of network performance, the model is an extremely helpful instrument. This is due to the fact that it is able to generate results that are not only clear but also simple to comprehend. The purpose of this thesis is to make use of ARIMA to generate short-term forecasts that are extremely accurate and precise. Such forecasts are necessary for proactive network management and optimization. The application of ARIMA makes it possible to forecast times when the network is experiencing high or low utilization. This, in turn, makes it possible to allocate resources in a more effective manner and enhances the quality of service (QoS) of the network as a whole within the network.

Within the context of this situation, ARIMA provides a few advantages. In addition to being able to effectively handle linear trends, it is also able to handle seasonality, which makes it suitable for prediction over a relatively short period of time. Additionally, ARIMA is founded on a tried-and-true methodology that produces results that are dependable and easy to understand. This is a significant advantage of the method. Since ARIMA can accurately and promptly predict network performance, it is a choice that is suitable for this project. Because of this, it has the potential to enhance the efficiency of 4G networks, which in turn increases the overall effectiveness of these networks.

The ARIMA model has three components that are required to help generate predictions. These parameters are estimated according to the dataset used, and help the model set the prediction pattern.

- *Autoregressive (AR) component:* The autoregressive part of ARIMA (denoted as 'p') indicates that the evolving variable of interest is regressed on its own prior values. For instance, in an AR(1) model, the current value is based on the immediately preceding value. In general, an AR(p) model uses the past 'p' values to predict the current value.

- *Integrated (I) component:* The integrated part (denoted as 'd') involves differentiating the data to make it stationary. A time series is stationary when its mean and variance are constant over time, and its covariance is time-invariant. Differencing is a transformation technique used to stabilize the mean of a time series by removing changes in the level of a time series, thereby eliminating trend and seasonality.
- *Moving Average (MA) component:* The moving average part (denoted as 'q') models the relationship between an observation and a residual error from a moving average model applied to lagged observations. Essentially, it smooths out the noise in the data, enabling better detection of the underlying signal.

These components collectively are referred to as ARIMA(p,d,q), where p, d and q stands for:

- 'p' is the order of the autoregressive part.
- 'd' is the degree of differencing required to achieve stationarity.
- 'q' is the order of the moving average part.

The parameters 'p' and 'q' are typically determined using the autocorrelation function (ACF) and partial autocorrelation function (PACF), while the values are estimated using maximum likelihood estimation (MLE).

ARIMA is utilized to forecast downlink throughput in 4G networks around Dhaka and Gazipur. The collected dataset underwent thorough preprocessing to ensure stationarity, which is a prerequisite for applying ARIMA. This preprocessing involved removing trends and stabilizing variance through differencing. Once the data was stationary, ARIMA models were fitted to capture the linear dependencies within the time series. The use of ARIMA in this project serves several purposes:

1) *Short-term forecasting:* ARIMA is particularly effective for short-term predictions, which are essential for real-time network management and optimization.
2) *Trend and seasonality detection:* By capturing linear trends and seasonal effects, ARIMA helps in understanding the temporal dynamics of network performance.

3) *Proactive network management:* The forecasts generated by ARIMA enable network operators to anticipate periods of high or low throughput, facilitating proactive measures to manage network resources efficiently

There are numerous positive aspects that contribute to making ARIMA an outstanding choice as a time series forecasting model. One notable advantage is its ability to effectively handle both linear trends and seasonality. ARIMA is highly effective in modeling the linear dependence between two datasets. This feature makes it well-suited for situations in which distinct trends and seasonal patterns are observable. This component ensures that the model can effectively represent and capture the inherent temporal patterns of the data. This enhances the credibility of the predictions.

The ARIMA model's capacity to provide unambiguous outcomes is one of its most notable attributes. As a network manager, it is crucial that the forecasts generated by ARIMA are concise and comprehensible. Network operators can make informed decisions and take appropriate actions when the predictions are easily comprehensible. This ensures that the actions are carefully planned and precisely targeted. Additionally, it is worth noting that ARIMA is a widely recognized and extensively researched method that has been employed for a considerable duration. In theory, it is founded on established and validated concepts, rendering it a robust framework for analyzing time series data. Individuals place their trust in it due to its extensive and renowned historical background. They are aware that it possesses great strength and reliability.

ARIMA can provide valuable insights into short-term forecasting. The system must possess the capability to generate precise short-term forecasts to enable network performance management to implement timely adjustments. By accurately predicting changes in network throughput, operators can effectively allocate resources, enhance performance, and deliver superior Quality of Service (QoS) to users. In rapidly evolving environments with high demand, such as telecommunications networks, proactive management plays a crucial role. Even marginal enhancements in the accuracy of predictions can yield significant impacts on operational outcomes. ARIMA is a highly effective method for analyzing time series data due to its independence from trends and seasonality, provision of clear and comprehensible results, reliance on a well-established approach, and proficiency in generating short-term forecasts. The utilization of this thesis demonstrates the efficacy of its application in enhancing predictive capabilities for network operations, thereby facilitating more efficient and superior network management strategies.

ARIMA is crucial in this subject as it offers a strong structure for predicting the short-term performance of 4G networks. The functionality of this tool enables the capture of linear trends and seasonal variations in the data, thereby enhancing the accuracy and reliability of forecasting. Additionally, it facilitates proactive network management strategies, optimizes resource allocation for users, and enhances quality of service (QoS). Although there are limitations associated with the use of nonlinear models, the advantages of ARIMA make it an indispensable tool for the specific requirements of this project.

Chapter 4: Model Results and Discussion

4.1. Evaluation of Throughput Predictive Models

The performance of three different machine learning algorithms, namely K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest, was evaluated by analyzing the model scores of each of these algorithms. The process of scoring a model involves methodically applying each model to a dataset and evaluating the accuracy of the predicted outcomes in comparison to the actual values. The purpose of this endeavor is to identify the algorithm that is most effective in recognizing the innate patterns that are present in the data and in producing the most precise forecasts.

Table 4.1: Machine Learning Model Comparison

Model	Score
K-Nearest Neighbors	56.62
Logistic Regression	56.02
Random Forest	71.27

As shown by the results in Table 4.1, KNN achieved a model score of 56.62, Logistic Regression scored 56.02, and Random Forest achieved a significantly higher score of 71.27. The results of the evaluation show that the performance of the Random Forest algorithm for calculating downlink throughput is better compared to the other two algorithms.

One possible explanation for the differences in outcomes is that the operational mechanisms and inherent characteristics of each algorithm are responsible for the variation. The K-Nearest Neighbors (KNN) algorithm is a non-parametric algorithm that assigns labels to data points by determining the label that is utilized by the several data points' subsequent neighbors. That is straightforward and uncomplicated. The performance of KNN is heavily influenced by the choice of distance metric and the number of neighbors (K), even though it can be effective for certain data sets. Additionally, when dealing with extensive datasets, KNN may experience difficulties due to the substantial computational resources required for calculating the distances between data points. As a result, this may cause KNN to encounter difficulties. The inadequate performance and an increased likelihood of overfitting can be the consequence of using a dataset that has high dimensionality, which is the case for predictive modelling.

Logistic regression is a parametric algorithm that, in contrast to another method, estimates the likelihood of a binary result by linearly combining input characteristics. This is done to determine the likelihood of the binary result. In situations where there is a relationship that is approximately linear between the features that are included in the input and the output, it is extremely useful for binary classification problems. Logistic regression therefore is a straightforward method, which is a disadvantage in this case because it might not be able to capture the complex and non-linear connections that exist within the data.

The Random Forest algorithm, on the other hand, succeeded in achieving a model score of 71.27, which was significantly higher than the scores that were achieved by the other two algorithms. It is possible to attribute better performance of this algorithm to some fundamental factors that are inherent to the Random Forest. Through utilization of bootstrap aggregation, the Random Forest algorithm generates different subsets of the training data and uses each of these subsets on a decision tree.

Regression is a statistical technique that involves calculating the average of the predictions made by each tree to arrive at the final prediction. In the process of Random Forest classification, the ultimate prediction is arrived at by selecting the prediction that has received the greatest number of votes from all the individual trees. As a result of the utilization of this ensemble approach, the variability in the model's forecasts is effectively reduced, which in turn alleviates the problem of overfitting, which is a prevalent concern when using individual decision trees.

The Random Forest algorithm ensures the diversity of trees and prevents any one feature from dominating the prediction process. It does this by selecting a subset of features at random at each split in the decision trees. This attribute enhances the model's ability to generalize effectively to data that is not familiar to it. It also enables the model to capture intricate relationships between features that may be missed by models that are simpler, such as Logistic Regression. Because it can effectively handle complex patterns and interactions among multiple network parameters, the algorithm is highly suitable for the network performance data that was used in this study. This is a result of its resistance to overfitting as well as its capacity to deal with data that contains a large number of dimensions.

In addition to Random Forest's capacity to generate estimations of feature importance, its ability to deal with missing data contributes to the enhancement of its interpretability and resilience. Researchers can improve their optimization efforts by acquiring knowledge about the fundamental factors that influence network performance. This is accomplished by gaining an understanding of the characteristics that have the greatest impact on predictions. When it comes to dealing with complex and high-dimensional data, the Random Forest algorithm's exceptional performance in this study demonstrates its robustness and strong generalization capability. Because of the ensemble nature of the model, it is possible to select subsets of data and features randomly. This helps to prevent overfitting and capture complex patterns in the data. For predictive modeling tasks that require complex and diverse datasets, such as the network performance data that was analyzed in this project, Random Forest is the best option to go with.

Systematically assembling and evaluating machine learning prediction models provided valuable insights into their ability to accurately assess network performance. Random Forest outperforms simpler models due to its ability to function effectively as a collective entity. In Dhaka, it is crucial to possess resilience due to rapid pace of change, high population density, and potential network disruptions. Before its utilization in the analysis, the material underwent meticulous cleaning and verification to ensure its uniformity. To enhance the quality of the data and model, various techniques were employed, including outlier removal, data augmentation, missing value imputation, and feature scaling. Subsequently, every algorithm underwent training and testing using the preprocessed dataset. The Random Forest algorithm proved to be the most suitable tool for the task due to its ability to effectively process large volumes of data and comprehensively comprehend its underlying patterns and characteristics.

Undoubtedly, selecting the appropriate machine learning models for the dataset and domain is of utmost importance. The Random Forest algorithm performed admirably on this project. This demonstrates the accurate prediction of the performance of a 4G network and paves the way for further investigation and advancement in this field. The findings of the study can be utilized by researchers, network operators, and policymakers to enhance the effectiveness of predictive models. Implementing this solution will enhance the dependability of networks and increase user satisfaction in metropolitan areas.

4.2. Evaluation of Time Series Forecasting Models

It is important to evaluate time series forecasting models based on specific metrics to help understand which model can create the most accurate and trustworthy forecasts of the downlink throughput. The Root Mean Squared Error (RMSE) was selected in this case as the evaluation metric. RMSE is a measurement that is widely used to determine the differences that exist between the values that are projected by a model and the values that are observed. The ability to significantly penalize large errors, which provides a clear indication of the predictive accuracy of the model, is the primary utility that it possesses. RMSE can be calculated using the following:

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

where $\hat{y}$ corresponds to the model predictions, y corresponds to the actual data sample values and N refers to the total number of data samples.

Table 4.2: Time Series Forecasting Model Comparison

Model	RMSE
ARIMA	0.447
FbProphet	3419.08

Table 4.2 shows the comparison between the two models selected for time series forecasting. During the investigation, it was necessary to take note of the fact that ARIMA model demonstrated a noteworthy reduction in the root mean square error (RMSE) value, which was at 0.447. RMSE value for FBProphet was significantly larger, coming in at 3419.08. It is possible to attribute the significant differences in performance to several important factors, which are frequently associated with the inherent qualities and prerequisites of each model.

The ARIMA model was developed specifically for the purpose of conducting short time series analysis. It is particularly well-suited for examination of datasets that display linear trends and seasonality. It is possible for the model to effectively model temporal dependencies because it can capture autoregressive patterns, apply difference in order to achieve stationarity, and make use of moving averages. The fact that the dataset was of a short-term nature and that there were linear trends that the model could take advantage of contributed to the fact that ARIMA was extremely successful in our situation. Autocorrelation and partial autocorrelation functions were utilized to determine the optimal sequence for the model's three components, which were the autoregressive (p), differencing (d), and moving average (q) components. This was necessary to achieve satisfactory tuning of the model. The ARIMA version was meticulously calibrated, which ensured that it was able to accurately capture the underlying patterns in the dataset, which ultimately led to a reduction in the root mean square error (RMSE).

The ARIMA model was developed specifically for the purpose of conducting short-term forecasting with comparatively high accuracy and reliability. It is particularly well-suited for examination of datasets that display linear trends and seasonality. It is possible for the model to effectively model temporal dependencies because it can capture autoregressive patterns, apply differencing to achieve stationarity, and make use of moving averages. The fact that the dataset was of a short-term nature and that there were linear trends that the model could take advantage of contributed to the fact that ARIMA was extremely successful in our situation. Autocorrelation and partial autocorrelation functions were utilized to determine the optimal sequence for the model's three components, which were the autoregressive (p), differencing (d), and moving average (q) components. This was necessary to achieve satisfactory tuning of the model. The ARIMA version was meticulously calibrated, which ensured that it was able to accurately capture the underlying patterns in the dataset, which ultimately led to a reduction in the root mean square error (RMSE). Figure 3.3 shows the prediction conducted by the ARIMA model from 9pm. The ARIMA model was able to accurately pick up the intricate patterns in the dataset and produce satisfactory results.

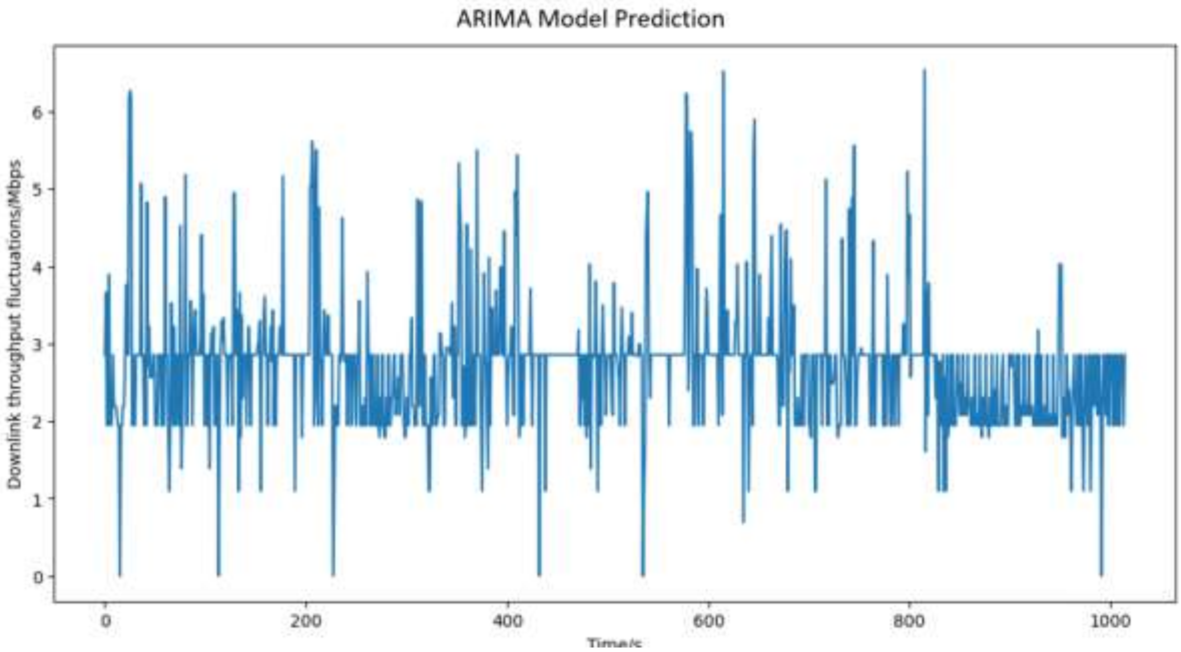

Figure 4.1: ARIMA model prediction

On the other hand, FBProphet is a model that has been developed with a specific collection of requirements and assumptions. The open-source tool known as FBProphet was developed in conjunction with the Data Science team at Facebook. The model was designed with the intent that the purpose of effectively managing large datasets that contain intricate seasonal patterns and holiday effects were satisfied. This model was found to be useful for long-term forecasting because of its efficiency where statistical data covers a period of several years. The ability of FBProphet to automatically model non-linear characteristics and seasonal patterns is one of its strengths. This ability also makes it user-friendly. In spite of this, one of the most significant drawbacks of this approach is that it necessitates the collection of a substantial amount of historical data, which typically spans a period of at least two years, in order to accurately capture and forecast long-term patterns and trends. Figure 3.2 illustrates the time series prediction done by using FBProphet over a 90-minute period. Certain portions of actual values of throughput were removed to see if the model can predict in the presence of null data. From the graph, it is clear that while it is able to make a prediction, it cannot model the actual fluctuations properly.

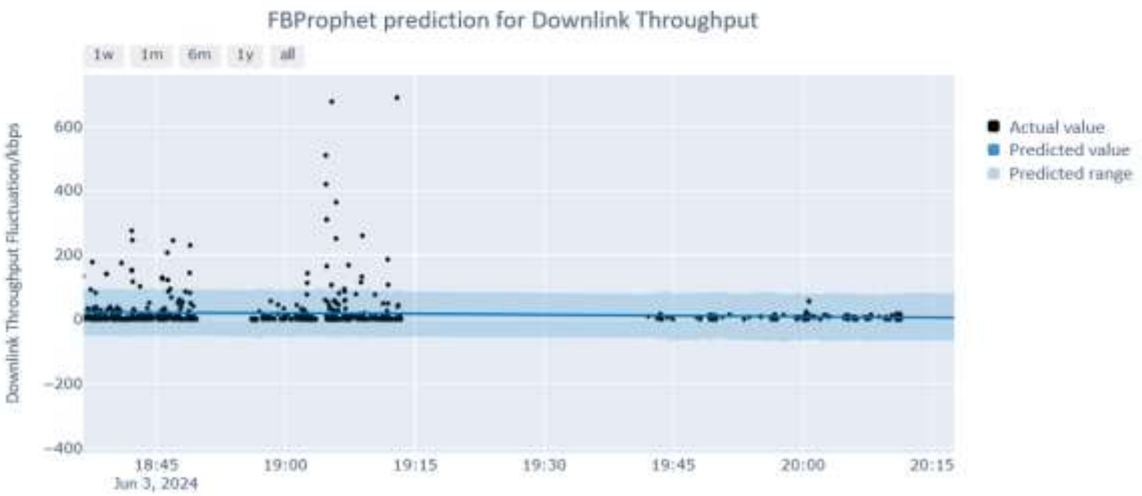

Figure 4.2: FBProphet prediction for Downlink Throughput

Over the course of the project, the dataset not grown to the extent where it would've been adequate to make full use of FBProphet's capabilities. It was difficult for FBProphet to accurately capture the long-term seasonal patterns that it was designed to model because the statistics only covered a short period of time. The disparity between the requirements of the model and the data that was available led to an increase in the number of prediction errors, which in turn led to a Root Mean Square Error (RMSE) that was significantly higher. The RMSE values that vary highlight how important it is to select the appropriate model solely based on the characteristics of the dataset and the time horizon that is being forecasted. When it comes to long-term forecasts that involve complex seasonal patterns, FBProphet is more effective than ARIMA. ARIMA is best suited for short-term predictions that involve linear trends. Therefore, it is essential for professionals to have a thorough understanding of these subtle distinctions to make educated decisions regarding which model to apply in specific circumstances.

As a result of the short-term characteristics and linear patterns of the dataset, the findings of this project ultimately demonstrate that ARIMA performed better than FBProphet in terms of root mean square error (RMSE), which draws attention to the fact that it is better suited for this particular dataset. On the other hand, the FBProphet model had a higher RMSE, and it needs a longer historical dataset to generate accurate forecasts. For achieving the highest possible level of predictive performance, this analysis highlights how important it is to align the model with the particular characteristics of the data.

Chapter 5: Demonstration of Outcome-Based Education (OBE)

A major paradigm shift in contemporary education is represented by Outcome-Based Education (OBE), which places an emphasis on the specific, quantifiable results that students should be able to get after finishing their courses. OBE places a higher priority on achieving competences, skills, and knowledge that students must demonstrate at the conclusion of their learning journey than traditional education models, which frequently place an emphasis on the delivery of content by instructors and the completion of mandated syllabi. This creative method is fundamentally student-centered, highlighting the vital requirement to customize instructional tactics and evaluation procedures to guarantee that every student meets specified, precisely defined learning objectives.

Learning outcomes are commonly expressed in terms of knowledge, abilities, attitudes, and values that students should gain within the framework of the OBE. These results can cover a wide range of areas, from interpersonal and ethical aspects to technical and cognitive abilities. Both teachers and students can have a common grasp of the learning objectives and success criteria thanks to the clarity and precision of these outcomes. Transparency encourages a feeling of shared accountability and community, which increases accountability and makes it possible for teaching and learning methods to continuously improve.

OBE assessments are intended to be genuine and performance-based, offering insightful information on students' capacity to apply what they have learned in real-world settings. These evaluation techniques go beyond the conventional tests and exams by including projects, portfolios, presentations, and other techniques that call on students to show their abilities in practical settings. This all-encompassing method of evaluation not only confirms that learning objectives are met, but it also improves students' critical thinking, problem-solving, and teamwork abilities.

Every stakeholder involved in the educational ecosystem—teachers, administrators, legislators, and students—must work together to implement OBE. It demands a change in perspective and methods, backed by continual professional development and a dedication to the constant assessment and improvement of instructional techniques. With an emphasis on innovation and continuous improvement, OBE seeks to develop a flexible and adaptable educational framework that equips students with the opportunities and challenges of the twenty-first century.

OBE has clear objectives that need to be addressed. These include the course and program outcomes, the knowledge profile required to complete the course, and attributes of the range of engineering activities and engineering problem solving associated with the course. Fulfilling each of the mentioned points ensures enhances learning experiences and skill development, which further emphasizes the student-based learning ideology of OBE.

5.1. Course Outcomes (COs) Addressed

The following table shows the COs addressed in EEE 4700 for Project and Thesis.

Table 5.1: List of course outcomes addressed in EEE 4700 for project and thesis

COs	CO Statement	POs	Put Tick (√)
			EEE 4700
CO1	Identify a contemporary real life problem related to electrical and electronic engineering by reviewing and analyzing existing research works.	PO2	√
CO2	Determine functional requirements of the problem considering feasibility and efficiency through analysis and synthesis of information.	PO4	√
CO3	Select a suitable solution and determine its method considering professional ethics, codes and standards.	PO8	√
CO4	Adopt modern engineering resources and tools for the solution of the problem.	PO5	√
CO5	Prepare management plan and budgetary implications for the solution of the problem.	PO11	√
CO6	Analyze the impact of the proposed solution on health, safety, culture and society.	PO6	√
CO7	Analyze the impact of the proposed solution on environment and sustainability.	PO7	
CO8	Develop a viable solution considering health, safety, cultural, societal and environmental aspects.	PO3	√
CO9	Work effectively as an individual and as a team member for the accomplishment of the solution.	PO9	√
CO10	Prepare various technical reports, design documentation, and deliver effective presentations for demonstration of the solution.	PO10	√
CO11	Recognize the need for continuing education and participation in professional societies and meetings.	PO12	√

The justification of the selected COs are as follows:

CO1: Identify a contemporary real-life problem related to electrical and electronic engineering by reviewing and analyzing existing research works

- *Real-life problem diagnosis:* Analyze the most recent research carefully to determine the specific problems that Bangladesh's 4G networks are facing. Identifying and addressing issues such as fluctuating changes in users' data speeds, network congestion-induced slowdowns during peak hours, and coverage gaps that result in little to no signal in some neighborhoods. To create solutions that will work, it is necessary to identify these specific difficulties.

- *Examining existing Machine Learning (ML) models*: Examine ML models that are currently in use for traffic predictions and network optimization. The difficulty lies in how these models might be modified to fit Bangladesh's particular circumstances. This involves considering variables that affect network performance, such as user mobility, regional variances, and even weather patterns. Make sure the models meet the problems that Bangladeshi 4G networks are facing by personalizing the models.

CO2: Determine functional requirements of the problem considering feasibility and efficiency through analysis and synthesis of information.

- *Feasibility check:* Analyze the chosen machine learning models in depth. Consider the necessary data when you train these models. Viability tests are performed to see whether it is practical to execute the models in an actual scenario. These checks include determining whether the relevant data is accessible, estimating the amount of processing power needed to train the model, and estimating the training duration. It is necessary to assess the models' efficiency as well.
- *User-centric functionalities:* To improve user experience, it is crucial to develop machine learning models with user-centric functionalities. These include predicting peak usage times, identifying congested areas, and suggesting optimal network configurations. Predicting peak usage times involves creating models that forecast network traffic peaks using historical and event-specific data, which improves resource allocation, service quality, and cost efficiency. Identifying congested areas entails deploying models that detect and predict network congestion through real-time usage and geographical data, enabling precise adjustments, enhancing reliability, and allowing proactive management. Suggesting optimal network configurations involves systems recommending tailored solutions based on local usage patterns and network topology, ensuring efficient performance, user satisfaction, and scalable resource management. The development process includes conducting user research to identify needs, collecting and analyzing relevant data, developing advanced forecasting, clustering, and recommendation algorithms, validating models through cross-validation and pilot tests, and iterating based on user feedback.

CO3: Select a suitable solution and determine its method considering professional ethics, codes and standards.

- *Ethical Machine Learning:* In today's data-driven world, ethical considerations are paramount. The training data used to build the ML model can harbor hidden biases. For instance, data collected primarily from urban areas might not accurately reflect network performance in rural regions. The task is to select ML techniques that mitigate these biases and ensure the models' fairness across all user demographics and locations. This might involve employing data augmentation techniques to create a more balanced training dataset and implementing fairness metrics to monitor and address potential biases during the development process.
- *Practices for the most accurate predictions:* Ensure the chosen model follows the best-established practices for ML prediction methodologies. This includes adhering to data pre-processing standards, a crucial step that cleanses and prepares the data for the model. Regularization techniques come into play to prevent the model from overfitting to the training data, ensuring it can generalize well to unseen scenarios. Implementing proper evaluation metrics is just as essential. These metrics assess the model's performance, allowing to refine and improve its accuracy over time.

CO4: Modern Tools for Powerful Solutions

- *Python:* The preferred language for creating and implementing a machine learning model is Python. One may easily develop complicated models with its vast ecosystem of libraries. There are many potent Python libraries that are very readily accessible for creating and refining machine learning models.
- *Data visualization:* Data visualization tools are crucial when trying to fully understand the variances in a dataset. They are critical to understanding network data patterns, assessing model results, and effectively communicating findings to both non-technical and technical audiences. Using libraries like Matplotlib and Seaborn, one may create informative graphics that show the model's effectiveness and probable impact on network optimization. These graphics can also be spatially plotted using programs like Google Earth.

CO5: Guiding the Project to Success: Management Plan and Budget

- *Project management plan:* Producing a roadmap that describes the whole project's path from the beginning to the end is essential to ensure the project in hand is carried out as efficiently and as effectively as possible. It lists important benchmarks, deadlines for every phase, and the resources necessary to reach them. It also defines clear communication strategies to ensure everyone involved is on the same page. The project's overall plan is for gathering enormous amounts of network performance data from particular interest regions, formatting the data appropriately, and producing a final dataset. The dataset is then fed into the machine learning model, and the model's predictions are assessed according to specified criteria.

- *Budget breakdown:* The project required certain costs to be executed. The project's budgetary requirements are broken down in detail below:
    - Personnel costs: Data acquisition involved travelling through the selected regions of interest at specific times of the day. To collect data, a network performance data logging software, namely G-NetTrack Pro, was used and the user had to travel across the different areas using different modes of transportation, both public and private. Multiple trips were required for collecting accurate and reliable data that was acceptable.
    - Hardware and software expenses: This includes the price of the network performance data recording program G-NetTrack Pro, which is the premium version of the G-NetTrack program, as well as the processing power required to train and implement the ML model.

CO6: Analyze the impact of the proposed solution on health, safety, culture and society.

- *Societal impact:* Though the implementation of 4G predictive models, the community as a whole experiences the benefits of better networking conditions. Digital inclusivity is achieved, as it allows for more individuals to make use of the higher standards of 4G technology and networks. As the importance of 4G technology grows in stature in Bangladesh, the development of job opportunities in the telecommunication sector also increases. Finally, due to better networking environment, 4G services become more economically accessible.

CO8: Develop a viable solution considering health, safety, cultural, societal and environmental aspects.

- *Safety aspect:* User safety should be paramount. Ensure the model doesn't emit harmful radiation or interfere with medical devices. Cultural sensitivity is also crucial. The model's functionalities should respect diverse cultures and avoid any potential biases. Network security of transmitting data over 4G networks should be of the utmost importance when designing systems based on predictive modelling. The systems should include measures to protect against data interception, hacking and other cyber threats.

CO9: Work effectively as an individual and as a team member for the accomplishment of the solution.

- *Literature review:* Researching papers and studies on 4G network optimization and ML techniques, sharing findings, and collectively building a strong foundation for the project.
- *Design and simulation:* Working collectively when planning out of the work of the project, and during the simulation of the different models. This ensures there is no communication gap between members and that everyone is actively involved throughout the project.
- *Presentation and documentation:* Meticulous work was put in to design presentations and preparing comprehensive reports. These reports detail findings of the project, the chosen methodology, and the results achieved, ensuring clear documentation for future reference.
- *Communication tools:* This point highlights the power of utilizing collaborative tools like Google Slides and Docs. These platforms allow everyone to work on documents simultaneously, fostering real-time updates and ensuring everyone is on the same page. Additionally, regular Google Meet meetings become the virtual space for brainstorming sessions, progress updates, and resolving any challenges that may arise.

CO10: Prepare various technical reports, design documentation, and deliver effective presentations for demonstration of the solution.

- *In-depth technical reports:* Creating reports that narrate the progress of the project, from start to finish. These reports also detail the research on 4G network optimization and ML techniques that were conducted in order to help understand the work-flow of the project. Additionally, they explain the methodology used to develop the solution, showcase the results achieved, including key performance metrics, and highlight the project's overall impact.
- *Documentation for future reference:* Designing and creating effective documents that serve as a roadmap for understanding the technical aspects of the project. The

documentations help explain the model's architecture, its functionalities, and the underlying algorithms in a way that is easy for individuals to grasp.

CO11: Recognize the need for continuing education and participation in professional societies and meetings.

- *Engaging with professional societies and attending meetings:* The platforms provided by professional societies and meetings offer a vibrant ecosystem for networking with other professionals in the fields of both Machine Learning as well as telecommunication. They provide valuable opportunities to exchange ideas, discuss best practices, and stay informed about the latest research and trends. By actively participating in these communities and potentially collaborating with others, it is possible to learn from their experiences, and stay at the forefront of technological advancements.

5.2. Aspects of Program Outcomes (POs) Addressed

The following table shows the aspects addressed for certain Program Outcomes (POs) in EEE 4700 for Project and Thesis.

Table 5.2: List of aspects addressed for certain program outcomes in EEE 4700 for projects and thesis

	Statement	**Different Aspects**	**Put Tick (√)**
PO3	**Design/development of solutions**: Design solutions for complex electrical and electronic engineering problems and design systems, components or processes that meet specified needs with appropriate consideration for public health and safety, cultural, societal, and environmental considerations.	Public health	
		Safety	√
		Cultural	
		Societal	
		Environmental	
PO4	**Investigation:** Conduct investigations of complex electrical and electronic engineering problems using research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of information to provide valid conclusions.	Design of experiments	√
		Analysis and interpretation of data	√
		Synthesis of information	√
PO6	**The engineer and society:** Apply reasoning informed by contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to professional engineering practice and solutions to complex electrical and electronic engineering problems.	Societal	√
		Health	
		Safety	
		Legal	
		Cultural	
PO7	**Environment and sustainability:** Understand and evaluate the sustainability and impact of	Societal	

	professional engineering work in the solution of complex electrical and electronic engineering problems in societal and environmental contexts.	Environmental	
PO8	**Ethics:** Apply ethical principles embedded with religious values, professional ethics and responsibilities, and norms of electrical and electronic engineering practice.	Religious values	
		Professional ethics and responsibilities	√
		Norms	
PO9	**Individual work and teamwork:** Function effectively as an individual, and as a member or leader in diverse teams and in multi-disciplinary settings.	Diverse teams	√
		Multi-disciplinary settings	√
PO10	**Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.	Comprehend and write effective reports	√
		Design documentation	√
		Make effective presentations	√
		Give and receive clear instructions	
PO11	**Project management and finance:** Demonstrate knowledge and understanding of engineering management principles and economic decision-making and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.	Engineering management principles	√
		Economic decision-making	√
		Manage projects	√
		Multidisciplinary environments	

The justification of the selected POs are as follows:

PO2: Problem Analysis

- *Data-driven network analysis:* The dataset is the key to understanding network behavior. Analyze factors influencing performance within the data, like cell tower traffic patterns, signal strength variations across locations, and user activity logs.
- *Identifying bottlenecks and opportunities:* Leverage the dataset to pinpoint areas experiencing bottlenecks or throughput limitations. These areas become targets for the model to optimize. Conversely, the analysis might reveal opportunities to enhance network efficiency in areas with underutilized resources based on usage patterns within the data.
- *Building context-specific understanding:* Develop a deep understanding of network behavior specific to the data. This context is crucial for selecting relevant data points and designing features that effectively capture the unique dynamics of the 4G network for the model.

PO3: Designing the Prediction Model

- *Model selection based on data characteristics:* Analyze the dataset's size, structure, and the types of features extracted to choose the most suitable ML models. This ensures the selected models can effectively learn from the data and predict network throughput within the specific network.

- *Data-driven model optimization:* The model hyperparameters are fine-tuned based on the characteristics of the data. This data-driven approach optimizes its ability to predict future network throughput within the specific network environment.

PO4: Investigate Network Data

- *Data exploration and feature selection:* Employ various data analysis techniques to explore the dataset. This might involve statistical methods to identify correlations between downlink throughput and signal strength, or data visualization tools to uncover hidden patterns in performance metrics over time. These insights become the foundation for developing features and ultimately, a powerful prediction model.

PO5: Modern Data Analysis Tool

- *Extensive use of various Python libraries:* Leveraging powerful data analysis software like Python libraries (e.g., Pandas, NumPy, Matplotlib) to clean and organize the dataset. This ensures the data is in a format suitable for training the model effectively [17-19].
- *Data preprocessing for accuracy:* Data quality is paramount! Implement data preprocessing techniques to ensure its accuracy and consistency. This might involve handling missing values, identifying and correcting outliers, and scaling the data to a common range. By cleaning and preparing the data, the model learns from reliable information, leading to more accurate predictions.

PO6: The Engineer and Network Access

- *Catalyzing economic growth:* The project can be considered as a driver of economic development. By optimizing network performance through ML, the model enables faster data speeds and improved reliability. This translates to a more attractive environment for businesses, potentially leading to job creation and economic growth, especially in sectors reliant on reliable internet connectivity.
- *Digital inclusion:* The improved network access in underdeveloped areas, facilitated by the prediction model, can bridge the digital divide. This provides opportunities for remote education, access to vital healthcare services, and the

growth of e-commerce, ultimately leading to an improved quality of life for communities.

PO8: Ethics and Data Privacy

- *Informed consent and transparency:* For real-time 4G predictive models to work on a large scale, user data needs to be collected. As a result, the users should be fully aware of this matter, and should be given a clear explanation behind the purpose of data collection. In addition, the data collection and analysis process should be kept transparent. This can be done through well-documented methodologies.
- *Minimization principle:* Only data that is required to accomplish the objectives of the project are collected. No data that might increase privacy risks is collected.

PO9: Individual Contribution and Teamwork

- *Maintaining team cohesion:* The project's success relies on a harmonious collaboration among team members. Being able to effectively communicate and work alongside members to ensure all aspects (data collection, model development, implementation) of the project are in perfect sync. This collaborative approach ensures a smooth transition from model development to real-world network optimization, just like a well-rehearsed orchestra seamlessly transitions between movements in a symphony.

PO10: Clear Communication of Findings

- *Data-driven problem solving:* Utilize the model's predictions to anticipate potential bottlenecks or areas of inefficiency. Address these issues by analyzing data patterns and implementing targeted network adjustments based on the model's insights.

PO11: Project Management and Finance

- *Budget management:* The project involved both hardware and software costs, as well as costs for collecting data by travelling through selected regions of interest in Gazipur and Dhaka. This required the careful budgeting of the project to ensure that none of the phases of the project exceeded the budget constraints.

5.3. Knowledge Profiles (K3 – K8) Addressed

The following table shows the Knowledge Profiles (K3 – K8) addressed in EEE 4700 for Project and Thesis.

Table 5.3: List of knowledge profiles (K3-K8) addressed in EEE 4700 for project and thesis

K	**Knowledge Profile (Attribute)**	**Put Tick (√)**
K3	A systematic, theory-based formulation of engineering fundamentals required in the engineering discipline	
K4	Engineering specialist knowledge that provides theoretical frameworks and bodies of knowledge for the accepted practice areas in the engineering discipline; much is at the forefront of the discipline	√
K5	Knowledge that supports engineering design in a practice area	√
K6	Knowledge of engineering practice (technology) in the practice areas in the engineering discipline	√
K7	Comprehension of the role of engineering in society and identified issues in engineering practice in the discipline: ethics and the engineer's professional responsibility to public safety; the impacts of engineering activity; economic, social, cultural, environmental and sustainability	√
K8	Engagement with selected knowledge in the research literature of the discipline	

The justification of the selected Knowledge Profiles are as follows:

K4: A systematic, theory-based formulation of engineering fundamentals required in the engineering discipline

- *Theoretical background:* The project requires a basic understanding of 4G networks and data analysis, and a strong background in data structures and machine learning. Understanding 4G networking is essential to know what network performance parameters are required to be collected and how each can affect downlink throughput. The knowledge of data structures and machine learning is required for preprocessing the dataset and creating the predictive models.

K5: Knowledge that supports engineering design in a practice area

- *Application of the project model in real life scenario:* Comprehensive data analysis methods and advanced predictive models provided by the project can be applied to improve network performance and design and thus helping telecommunications engineering decisions in turn.

K6: Knowledge of engineering practice (technology) in the practice areas in the engineering discipline

- *Application of the model by network operators:* Through integrating real-world scenarios with machine learning and time series forecasting technologies, it was possible to show how these methods can be applied to improve 4G network performance and throughput forecasts in the telecommunication industry.

K7: Comprehension of the role of engineering in society and identified issues in engineering practice in the discipline: ethics and the engineer's professional responsibility to public safety; the impacts of engineering activity; economic, social, cultural, environmental and sustainability

- *Minimizing the digital divide:* The project aims to help alleviate the problem of digital divide by attempting to solve issues like guaranteeing public safety by dependable network performance, realizing the financial benefits of network efficiency optimization, and considering the social, cultural, and environmental effects of the development of telecommunications infrastructure.

5.4. Attributes of Ranges of Complex Engineering Problem Solving (P1 – P7) Addressed

The following table shows the attributes of ranges of Complex Engineering Problem Solving (P1 – P7) addressed in EEE 4700 for Project and Thesis.

Table 5.4: List of attributes of ranges of complex engineering problem solving (P1-p7) addressed in EEE 4700 for project and thesis

P	**Range of Complex Engineering Problem Solving**	**Put Tick (√)**
Attribute	Complex Engineering Problems have characteristic P1 and some or all of P2 to P7:	
Depth of knowledge required	**P1:** Cannot be resolved without in-depth engineering knowledge at the level of one or more of K3, K4, K5, K6 or K8 which allows a fundamentals-based, first principles analytical approach	√
Range of conflicting requirements	**P2:** Involve wide-ranging or conflicting technical, engineering and other issues	√
Depth of analysis required	**P3:** Have no obvious solution and require abstract thinking, originality in analysis to formulate suitable models	√
Familiarity of issues	**P4:** Involve infrequently encountered issues	
Extent of applicable codes	**P5:** Are outside problems encompassed by standards and codes of practice for professional engineering	
Extent of stakeholder involvement and conflicting requirements	**P6:** Involve diverse groups of stakeholders with widely varying needs	
Interdependence	**P7:** Are high level problems including many component parts or sub-problems	

The justification of the selected attributes of complex engineering problem solving are as follows:

P1: Depth of knowledge required

- *Comprehensive literature review:* Extensive review of related work and studies about 4G network optimization, data analysis and predictive models. Allowed for a clear idea and work plan of how to conduct the project as efficiently and effectively as possible.
- *Theoretical foundation:* Topics related to 4G telecommunication and network parameters needed to be studied carefully to comprehend the project at hand. The knowledge of data structures and machine learning is required to preprocess the obtained dataset so that they work as effectively as possible, and to implement the predictive models that provide reliable and accurate results.
- *Technical proficiency:* The project requires high proficiency in Python, Microsoft Excel as well as SQL. Excel and SQL are required to set the dataset in the correct format, whereas Python is required to analyze 4G data, and generate the throughput prediction models.

P2: Range of conflicting requirements

- *Scalability and complexity:* As the model grows and is implemented on a larger scale, the overall dataset complexity increases. There hence is a tradeoff between how large it is possible to scale the dataset of the project and the complexities involved in the dataset, such as more feature variables that influence the final throughput prediction.

P3: Depth of analysis

- *Detailed data analysis:* The finalized dataset was analyzed in order to comprehend the underlying patterns or trends in the feature variables, and observe how changes in the variables affect throughput, both through correlation analysis and by visualization tools such as Google Maps and the Matplotlib library of Python.
- *Model validation and comparative analysis:* The models used in the project were evaluated based on selected evaluation metrics. The evaluation of the different models was then used to compare different models to each other and observe which of the selected models had the best performance.

5.5 Attributes of Ranges of Complex Engineering Activities (A1 – A5) Addressed

The following table shows the attributes of ranges of Complex Engineering Activities (A1 – A5) addressed in EEE 4700 for Project and Thesis.

Table 5.5: List of attributes of ranges of complex engineering activities (A1–A5) addressed in EEE 4700 for project and thesis

A	**Range of Complex Engineering Activities**	**Put Tick (√)**
Attribute	Complex activities mean (engineering) activities or projects that have some or all of the following characteristics:	
Range of resources	**A1:** Involve the use of diverse resources (and for this purpose resources include people, money, equipment, materials, information and technologies)	√
Level of interaction	**A2:** Require resolution of significant problems arising from interactions between wide-ranging or conflicting technical, engineering or other issues	√
Innovation	**A3:** Involve creative use of engineering principles and research-based knowledge in novel ways	√
Consequences for society and the environment	**A4:** Have significant consequences in a range of contexts, characterized by difficulty of prediction and mitigation	
Familiarity	**A5:** Can extend beyond previous experiences by applying principles-based approaches	

The justification of the selected attributes of complex engineering activities are as follows:

A1: Range of resources

- *Literature and academic resources:* The project required thorough review of relevant scholarly articles, books and conference papers.
- *Software and tools:* Specialized software such as G-Nettrack, Python and SQL were utilized to complete the project.

A2: Level of interaction

- *Academic collaboration:* Worked alongside academic advisors and faculty members in order to refine project methodology and modelling for downlink throughput.

A3: Innovation

- *Novelty of aataset:* The dataset used in the project was all collected by members of the project, using the G-Nettrack software. This involves travelling across different selected regions of Gazipur and Dhaka on different modes of transport at specified time intervals.

Chapter 6: Future Scopes

As we consider the future, there are numerous intriguing possibilities for expanding and enhancing the predictive modeling research conducted in this thesis. Each of these guidelines has the potential to significantly improve the accuracy and implementation of community performance predictions, particularly as we move towards the adoption of advanced mobile technologies such as 5G and 6G.

- ***Combination with 5G/6G networks***

The implementation of 5G technology and the subsequent development of 6G networks have the potential to bring about changes in the field of telecommunications that have never been seen before. When compared to 4G networks, these next-generation networks offer improved data rates, decreased latency, and increased capacity. In order to extend our predictive models to include 5G and 6G networks, we will need to make adjustments to novel network structures. These structures include network slicing and edge computing, both of which are essential for these technologies. Our models take advantage of the particular characteristics of 5G and 6G networks, such as millimeter-wave frequencies and large MIMO, in order to provide accurate predictions that are tailored to the more advanced capabilities of these networks.

- ***Deep learning power***

The utilization of advanced deep learning methodologies, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), has the potential to significantly improve the accuracy of network performance evaluations [20]. Because of their ability to store and make use of information from earlier time steps, Recurrent Neural Networks (RNNs) are an excellent choice for processing time series data. Because of this, they are very good at capturing the temporal dependencies that are present in network performance metrics. Conversely, Convolutional Neural Networks (CNNs) can extract spatial features from gridded data, thereby facilitating the analysis of the impact of geoenvironmental factors.

- ***Context matters***

When contextual factors like weather conditions, individual behavior, and specific network events are incorporated into predictive models, the flexibility and precision of these models can be significantly improved. Weather conditions can have an effect on the propagation of signals, individual behavior patterns can have an effect on the load on a community, and events like concerts or sporting events can cause unexpected surges in network usage. Models have the potential to become more dynamic and adaptable to situations that occur in the real world if these variables are incorporated. It is guaranteed by this all-encompassing approach that predictions are not only derived from the performance of the network in the past, but that they also consider external factors that have the potential to influence the behavior of the network.

- ***Real time edge computing***

The optimization of real-time networks is made possible through the combination of predictive models and edge computing. The term "edge computing" refers to the practice of positioning computing resources near the source of the data. This results in a reduction in latency and makes it possible to accomplish data processing more quickly. When network operators deploy predictive models at the edge of the network, they can make immediate modifications to the architecture of the network configuration. This method is particularly useful for applications that are relatively straightforward, such as virtual reality and autonomous vehicles, because it improves both the performance of these applications and the user experience.

- ***Personal throughput***

It is possible to predict the throughput of male and female users by analyzing their usage patterns and locations. This allows for a more customized network experience to be provided to the users. By analyzing the statistics that are gathered from individual consumer devices, models are able to accurately predict the expected performance of the network for each individual user. With the help of this individualized approach, network resources can be managed more effectively, ensuring that customers receive optimal performance that is tailored to their particular needs and circumstances. Furthermore, it has the capability to enable targeted interventions, such as the provision of additional bandwidth to users who are experiencing inadequate connectivity.

- ***Additional future scopes***

Integration with IoT devices: As the number of Internet of Things (IoT) devices continues to grow, incorporating data from these devices can provide additional insights into network performance. IoT devices can act as sensors, providing real-time data on network conditions, which can be used to enhance predictive models.

Cross-domain applications: Extending predictive models to other domains such as smart cities, healthcare, and industrial automation can provide broader applications of the technology. For instance, predicting network performance in smart city infrastructure can help in managing traffic flow and public safety.

Adaptive learning models: Developing adaptive learning models that can update themselves in real-time based on new data can further enhance the accuracy and relevance of predictions. These models can continuously learn from the latest network performance data, ensuring that they remain up-to-date with the current network conditions.

Enhanced privacy and security: As predictive models become more sophisticated, ensuring the privacy and security of the data used becomes increasingly important. Developing methods to anonymize data while maintaining its utility for prediction can help in addressing privacy concerns.

In summary, the trajectory of the future of predictive modeling for network performance is expansive and full of promise, presenting a multitude of opportunities for expansion and inventiveness. Through the utilization of cutting-edge technologies such as 5G and 6G networks, deep learning, and edge computing, we are able to improve the accuracy and efficiency of our prediction models. Because of this technology, it is possible to analyze complex data models and provide instantaneous insights, both of which are essential for ensuring that the performance of the network remains at its highest level in environments that are constantly shifting and posing challenges. Through the utilization of this all-encompassing approach, we ensure that our forecast is not solely dependent on historical data, but rather takes into account the immediate external factors that have an effect on the performance of the network. These enhancements will not only improve the user experience and optimize the performance of the network, but they will also create opportunities for innovation in the field of telecommunications research and development. As we continue to develop and expand these predictive models and applications, their potential will expand beyond the realm of telecommunications to encompass a variety of other industries, including smart cities, healthcare, and industrial automation, among others. It is anticipated that these innovative technologies will bring about a revolution in these fields. As a consequence of the merger, there will be significant expansion and innovation in the fields of network performance, predictive management and advanced data analytics techniques.

Chapter 7: Conclusion

The project we carried out comprised a thorough investigation and the accumulation of extensive knowledge in the analysis of 4G data and the development of models for predicting downlink throughput. Our approach involved incorporating a wide range of advanced and creative techniques. Following the meticulous collection of data, we utilized tools to preprocess it, ensuring the stability and dependability of our dataset. This allowed us to comprehensively capture various network performance aspects in different urban settings. Our in-depth exploration of machine learning models, including K-nearest neighbor and random forest, showcased the broad range of applications and capabilities of these algorithms. Ultimately, we highlighted the heightened effectiveness of cluster methods in producing reliable and widely applicable predictions from data models.

The significance of choosing appropriate tools for various data types and forecasting needs was made clear by employing time series forecasting models such as ARIMA and FBProphet. While additional data is necessary, the straightforwardness and capacity of FBProphet to integrate external events provided our analysis with a new and insightful viewpoint. The accuracy of ARIMA and its capacity to produce near-term forecasts utilizing linear trends were extraordinary. The literature review provided a thorough investigation of previous research, covering past context and comparative evaluation. Our work showcased how it expands and improves upon previous research efforts. Through the implementation of a comprehensive approach, we successfully identified the distinct advantages and disadvantages of different models and methodologies, leading to a deeper understanding of their practical uses. The ability to integrate advanced technologies such as 5G and 6G networks, edge computing, and deep learning algorithms allows for the exploration of new possibilities in real-time communication and prediction accuracy within a high-quality system, with an emphasis on the future. Personalization is improved by taking into account the unique needs and usage patterns of individual users when predicting their throughput.

In summary, this extensive study not only showcases the impressive progress made in analyzing and forecasting 4G networks, but it also lays a firm basis for future research. By employing state-of-the-art tools and methodologies, we can push the boundaries even further, resulting in improved network efficiency, heightened user satisfaction, and a more interconnected global society.

References:

1. Shams AB. Abied SR. Hoque MA. Impact of user mobility on the performance of downlink resource scheduling in Heterogeneous LTE cellular networks, *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, Dhaka, Bangladesh, 2016, pp. 1-6, doi: 10.1109/CEEICT.2016.7873091.
2. Ira FLA, Ali Nurdin, Suroso. An Analysis of 4G Lite Signal Quality on Telkomsel XL Provider and Hutchinson 3 Indonesia Using G-Nettrack Pro Application Via Android at State Polytechnic of Sriwijaya. E-Komtek [Internet]. 2022 Dec.31;6(2):215-226. Available from: https://jurnal.politeknik-kebumen.ac.id/E-KOMTEK/article/view/958.
3. El-Saleh AA. Alhammadi A. Shayea I. Hassan WH. Honnurvali MS. Daradkeh YI. Measurement analysis and performance evaluation of mobile broadband cellular networks in a populated city. *Alexandria Engineering Journal*. 2023;66(1):927-946. doi: 10.1016/j.aej.2022.10.052
4. Ekeocha AC. Elechi P. Nosiri OC. Performance Analysis of KPI's of a 4G Network in a Selected Area of Port Harcourt, Nigeria. *World Journal of Electrical and Electronic Engineering*. 2021;1(1):44-50. Available from: https://www.scipublications.com/journal/index.php/wjeee/article/view/133 (accessed on 15 June 2024)
5. Elsherbiny H. Abbas HM. Abou-zeid H. Hassanein HS. Noureldin A. 4G LTE Network Throughput Modelling and Prediction. *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*. Taipei, Taiwan, 2020, pp. 1-6, doi: 10.1109/GLOBECOM42002.2020.9322410.
6. H. TK. The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no.8, pp. 832–844, 1998.
7. Kamakaris T. Nickerson J.V. Connectivity maps: Measurements and applications. *38th Annual Hawaii International Conf. on System Sciences, Big Island, HI, USA*. pp. 307–307.
8. Pögel T. Wolf L. Prediction of 3G network characteristics for adaptive vehicular connectivity maps. *Vehicular Networking Conf. (VNC), IEEE,* 2012. pp. 121–128.

9. Pögel T. Wolf L. Optimization of vehicular applications and communication properties with connectivity maps. *Local Computer Networks Conference Workshops (LCN Workshops), IEEE 40th*, 2015. pp. 870–877.
10. Xu Q. Mehrotra S. Mao Z. J. Li. PROTEUS: network performance forecast for real-time, interactive mobile applications. *11th Annual International Conf. on Mobile Systems, Applications, and Services*, 2013. pp. 347–360.
11. Liu Y. Lee J.Y. An empirical study of throughput prediction in mobile data networks. *Global Communications Conference (GLOBECOM), IEEE*, 2015. pp. 1–6.
12. Jin R. Enhancing upper-level performance from below: Performance measurement and optimization in LTE networks. Doctoral Dissertations, Univ. of Connecticut, United States, 2015.
13. Samba A. Busnel Y. Blanc A. Dooze P. Simon G. Instantaneous Throughput Prediction in Cellular Networks: Which Information Is Needed? *IFIP/IEEE International Symposium on Integrated Network Management (IM)*, Lisbonne, Portugal, 2017.
14. Elsherbiny, A. M. Nagib, and H. S. Hassanein, "4G LTE User Equipment Measurements along Kingston Transit 502 Bus Route," 2020. [Online]. https://doi.org/10.5683/SP2/EQWKO1 (accessed on 17 June 2024)
15. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion and O. Grisel, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, pp. 2825–2830, 2011.
16. Hyndman R.J. Athanasopoulos G. Forecasting: Principles and practice. Vic. Heathmont: OTexts, 2014.
17. Oliphant T.E. A guide to NumPy (Vol. 1). Trelgol Publishing, USA, 2006.
18. McKinney W. Data structures for statistical computing in python. *9th Python in Science Conf.*, vol. 445, pp. 51–56., 2010.
19. Hunter J.D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, vol. 9, no.3, pp. 90–95, May-June 2007.
20. Dupond S. A thorough review on the current advance of neural network structures. *Annual Reviews in Control*, vol. 14, pp. 200–230, 2019.