

مصعب السلامه

Musab Alsalamah v4.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:375870237

Submission Date

Apr 13, 2026, 3:05 AM GMT+2

Download Date

Apr 16, 2026, 1:34 AM GMT+2

File Name

Musab Alsalamah v4.docx

File Size

269.3 KB

39 Pages

6,982 Words

43,738 Characters

94% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups



58 AI-generated only 90%

Likely AI-generated text from a large-language model.



4 AI-generated text that was AI-paraphrased 5%

Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.





Predicting and Mitigating Urban Traffic Congestion Using Artificial Intelligence

Prepared By

Name: [Musab Abdullah Alsalamah]

ID No: [252000152]

Submitted in fulfillment of the requirements for the degree Master of
Science in Modern artificial intelligence

At

College of Informatics

Midocean University

UAE

Supervisor Name: Dr. Yasser Abdel Hamid

College of Informatics, Midocean University, Comoros

[2025-3]

Abstract

Traffic congestion is still a major challenge in city settings, particularly at signalized intersections where vehicle queues can build up rapidly during peak periods. This study suggests a machine learning-based approach for predicting short-term traffic congestion as well as examining its possible use in responsive traffic signal control.

In this study, I used traffic volume data collected from kaggle.com. After processing the data and extracting applicable features, I evaluated several machine learning models, including logistic regression, random forest, support vector machine, and gradient enhancement. Among these, the Gradient Boosting model attained the best performance, with an accuracy of 0.968, precision of 0.893, recall of 0.977, and F1-score of 0.933.

In addition to prediction, a simplified responsive traffic signal mechanism was implemented to simulate how predicted congestion can be used to adjust signal timing. The simulation results suggest that increasing green signal duration during predicted congestion periods may help reduce potential vehicle accumulation and boost traffic flow.

The results show that machine learning methods can effectively forecast traffic delays and support decision-making in traffic management. While the adaptive control was evaluated in a simplified simulation environment, the findings show the capability of integrating prediction models with traffic signal strategies for improving urban traffic performance.

Table of Contents

Abstract.....	2
Table of Contents	3
List of Abbreviations.....	5
List of Figures.....	6
List of Tables.....	7
Chapter 1: Introduction	8
1.1 Problem Definition.....	8
1.2 Thesis Objectives.....	9
1.3 Thesis Contributions.....	10
1.4 Thesis Organization	10
Chapter 2 Background and Related Work	12
2.1 Introduction.....	12
2.2 Classical Machine Learning Methods.....	12
2.3 Deep Learning Approaches.....	13
2.4 Hybrid and Optimized Models	14
2.6 Research Gaps and Opportunities	15
2.8 Previous Studies on Traffic Congestion Forecasting	17
2.9 Comparative Summary Table.....	18
2.10 Summary.....	19
Chapter 3 Proposed Approach.....	21
3.1 Overview	21
3.3 Data Preprocessing	22
3.4 Feature Engineering	23
3.5 Model Selection	24
3.6 Training Process	26
3.7 Model Evaluation	27
3.8 Adaptive Traffic Signal Mechanism	28

3.9 Summary.....	28
Chapter 4: Experimental Results and Discussion	30
4.1 Overview	30
4.2 Model Performance	30
4.3 Analysis of Predictions.....	31
4.4 Adaptive Traffic Signal Evaluation	32
4.5 Discussion.....	34
4.6 Summary.....	35
Chapter 5: Conclusion and Future Work	36
5.1 Introduction	36
5.2 Future Work	37
References	38

List of Abbreviations

- **AI:** Artificial Intelligence
- **ANN:** Artificial Neural Network
- **ARIMA:** Autoregressive Integrated Moving Average
- **BiLSTM:** Bidirectional Long Short-Term Memory
- **CNN:** Convolutional Neural Network
- **CPO:** Crested Porcupine Optimizer
- **DT:** Decision Tree
- **DeepONet:** Deep Operator Network
- **F1-score:** F1 Measure
- **GB:** Gradient Boosting
- **GAT:** Graph Attention Network
- **GRU:** Gated Recurrent Unit
- **IoT:** Internet of Things
- **ITS:** Intelligent Transportation Systems
- **KNN:** k-Nearest Neighbors
- **LightGBM:** Light Gradient Boosting Machine
- **LSTM:** Long Short-Term Memory
- **MAE:** Mean Absolute Error
- **MAPE:** Mean Absolute Percentage Error
- **ML:** Machine Learning
- **MLP:** Multi-Layer Perceptron
- **PeMS:** Performance Measurement System
- **PI-DeepONet:** Physics-Informed Deep Operator Network
- **R²:** Coefficient of Determination
- **RF:** Random Forest
- **RMSE:** Root Mean Square Error
- **RNN:** Recurrent Neural Network
- **SHAP:** SHapley Additive exPlanations
- **STGCN:** Spatio-Temporal Graph Convolutional Network
- **SVM:** Support Vector Machine
- **XAI:** Explainable Artificial Intelligence
- **XGBoost:** Extreme Gradient Boosting

List of Figures

Figures 1 22

Figures 2 24

Figures 3 26

Figures 4 26

Figures 5 27

Figures 6 28

Figures 7 31

Figures 8 33

List of Tables

Comparative Table 1 19

Chapter 1: Introduction

1.1 Problem Definition

Traffic jams at intersections with traffic lights are most obvious during rush hour. Drivers end up waiting longer, using more fuel, and creating more emissions. Most signal systems, such as fixed-time and semi-actuated types, use preset timing yet cannot respond to current traffic. As a result, they may fail to handle sudden increases in traffic demand, which leads to queue formation, longer travel times, and additional environmental impact.

Although congestion trends tend to repeat during peak hours, identifying exactly which traffic approaches will experience excessive demand continues to be a complex task. Traditional traffic analysis uses historical data or fixed rules, but these approaches cannot spot complex patterns or changes as they happen. As a result, congestion is usually noticed after it starts, which makes it harder to manage effectively (Medina-Salgado et al., 2022).

At the same time, the growing availability of traffic data from detectors and intelligent transportation systems provides an opportunity to improve congestion management. By applying artificial intelligence plus machine learning techniques, it becomes possible to analyze traffic behavior more effectively and predict congestion before it occurs. This study centers on the need for better ways to predict congestion at signalized intersections. Improving this area is important for developing traffic signal control systems that can adjust to traffic conditions and help intersections work more efficiently. However, although traffic congestion prediction has been widely studied, most research still focuses on refining model accuracy without linking the results to real traffic signal control systems. In practice, congestion is often predicted for analysis purposes only, rather than being used to support real-time or forward-looking decision-making.

Many recent studies use complex deep learning models that need a lot of computing power. As a result, these models are often not practical for actual traffic control.

In contrast, this study suggests a simplified yet effective framework which does not only predicts short-term congestion using structured traffic data, but also directly translates these predictions into responsive traffic signal control decisions. The novelty of this work consists in linking prediction outputs with implementable signal timing strategies, allowing proactive traffic management rather than reactive responses.

1.2 Thesis Objectives

This study is guided by the following question:

Can a Gradient Boosting model outperform a simple historical average method in predicting short-term traffic congestion at urban intersections?

Hypothesis: The main hypothesis of this study is that a Gradient Boosting model can achieve higher prediction quality than a baseline model based on historical averages in predicting short-term traffic congestion.

Additionally, it is hypothesized that uniting prediction results with responsive traffic signal timing can improve traffic movement by lowering congestion buildup at intersections. In addition, the study considers the following sub-question:

Can predicted congestion levels be effectively used to support responsive traffic signal timing decisions that reduce potential vehicle accumulation?

To achieve this objective, the study concentrates on several specific goals. First, it aims to analyze real-world traffic detector data in order to identify temporal patterns and traffic movements associated with congestion formation.

Apprehending these patterns is essential for building reliable predictive systems.

Second, I examine several machine learning and deep learning techniques that can be used for short-term congestion prediction in urban traffic environments. The goal is to evaluate how well these models capture traffic patterns.

Third, I apply feature engineering and feature selection techniques to improve model performance while maintaining interpretability. Focusing on the most relevant features helps improve prediction accuracy and gives better insight into the factors that influence traffic congestion.

In addition, the study appraises the performance of the developed models using standard classification metrics including accuracy, precision, recall, and F1-score. This secures a comprehensive assessment of the models' ability to detect congestion effectively.

Finally, the study intends to provide analytical and visual insights which can support proactive traffic management decisions. These results are intended to assist traffic operators toward refining signal control strategies and advancing overall intersection performance.

Overall, these objectives add to the wider goal of improving urban mobility via data-driven and intelligent transportation systems.

1.3 Thesis Contributions

This thesis introduces a clear and repeatable system for predicting traffic congestion at signalized intersections using real detector data. The framework tracks short-term congestion trends at each approach, providing a practical tool for smarter traffic evaluation and decision making.

A key contribution of this study lies in integrating congestion prediction with traffic signal control concepts. Instead of treating congestion forecasting as a separate analytical task, this approach shows how prediction results can be used to support adaptive signal timing strategies. In particular, the framework allows early adjustments to green-time allocation for heavily loaded traffic movements, which helps reduce queue buildup and increase overall intersection performance.

This work brings together recent developments in machine learning, deep learning, and combined models for traffic congestion forecasting. By applying these methods to real traffic data, it delivers a practical springboard for future work in intelligent transportation systems.

The thesis also points out the need to connect predictive analytics with day-to-day traffic management. This link makes the framework more useful and supports its use inside real-time systems, network-wide coordination, and advanced control methods like reinforcement learning for traffic signal optimization.

1.4 Thesis Organization

This thesis is organized into five main chapters.

Chapter 1 introduces the research background, defines the problem, outlines the objectives, and spotlights the main contributions of the study.

Chapter 2 introduces the theoretical background and discusses previous work on traffic congestion prediction, including classical machine learning methods, deep learning approaches, and hybrid models.

Chapter 3 describes the suggested methodology, including dataset description, data preprocessing, feature engineering, model selection, training process, and evaluation, as well as the responsive traffic signal mechanism.

Chapter 4 presents the experimental results and discusses the model effectiveness, prediction analysis, and the effectiveness of the adaptive signal system.

Chapter 5 concludes the study and provides recommendations for future work.

Chapter 2 Background and Related Work

2.1 Introduction

Traffic congestion forecasting has become an essential element of modern intelligent transportation systems (Medina-Salgado et al., 2022; Rahmani et al., 2023). especially since metropolitan regions continue to experience increasing traffic demand. At signalized intersections, traditional traffic management approaches, such as fixed signal plans based on historical averages, often struggle to handle short-term changes in traffic flow. Because of this, it becomes more difficult for traffic authorities to manage congestion in an effective and responsive way.

Recent improvements in data collection, sensing technology, and computing power have made it easier to use machine learning and deep learning for traffic analysis. These data-driven methods can find complex patterns and nonlinear relationships in traffic data better than established approaches. helps improve the accuracy and flexibility of congestion prediction models.

In addition to methodological developments, a growing body of research has studied different modeling methods for traffic congestion forecasting, ranging from classical machine learning methods to advanced deep learning and combined models. These studies give important knowledge of the advantages and disadvantages of each approach, as well as their usefulness for actual traffic contexts.

AI-based congestion forecasting is progressively regarded as an important tool for responsive traffic signal control and smart city projects. By predicting congestion in advance, these systems help traffic managers take action early, which improves how intersections work and keeps traffic moving smoothly.

This chapter provides the theoretical background and reviews the main literature on traffic congestion forecasting. It looks at current methods, points out important developments, and identifies research gaps that support the approach used in this study.

2.2 Classical Machine Learning Methods

Early studies in traffic congestion prediction relied mainly on classical machine learning techniques applied to structured traffic datasets. These methods were

primarily used to classify congestion levels or estimate short-term traffic conditions based relying on historical data and basic temporal features (Vlahogianni et al., 2014).

Common algorithms in this category include logistic regression, decision trees, support vector machines (SVM), k nearest neighbors (KNN), and ensemble algorithms such as the Gradient Boosting classifier. Gradient boosting builds models sequentially, where each new model focuses on correcting the errors of the previous ones, causing better predictive performance and better management of complicated data patterns.

Among these methods, tree-based models, particularly decision trees and random forests, received notable attention because of to their sturdiness and explainability. Random forests, for example, improve prediction performance by combining multiple decision trees, which helps reduce overfitting as well as enhances generalization when dealing with noisy traffic data.

Despite their advantages, classical machine learning models have considerable limitations. In particular, they have difficulty capturing complicated time correlations and nonlinear relations between traffic flows among different intersections. This limitation reduces their effectiveness in highly rapidly changing urban traffic environments, in which traffic patterns can change rapidly over time.

2.3 Deep Learning Approaches

Deep learning has refined traffic congestion forecasting because these models can learn complex, time-based patterns directly from traffic data. They usually work better than traditional machine learning methods for handling sequences and finding patterns through time in traffic flow (Kashyap et al., 2022).

Recurrent neural networks (RNNs), especially Long Short-Term Memory (LSTM) models, are commonly used for traffic prediction. They handle time-series data well and can spot long-term patterns. LSTM models usually predict short-term congestion more accurately than classic methods. In addition, attention methods have been introduced to boost model effectiveness by focusing on the most relevant temporal features (Fang et al., 2022).

Hybrid deep learning models have also been developed to further improve forecasting performance. For example, CNN–LSTM models combine convolutional units to detect spatial attributes with recurrent layers to learn time-

based patterns. This combination helps represent both local traffic changes and overall traffic trends.

More recently, graph-based deep learning models have gained attention in traffic forecasting research. Physics-informed models have also been introduced to incorporate domain knowledge into traffic prediction, improving both exactness and robustness (Wang et al., 2021). By representing traffic networks as graphs, these models can represent spatial connections between intersections while also modeling temporal changes.

These spatiotemporal approaches provide a better understanding of how congestion spreads over city road networks. Even though deep learning models perform well, they often need a lot of data, longer training times, and powerful computers. This makes them hard to use in real road systems, especially where infrastructure is limited or quick responses are needed. As a result, there is still a need for increasingly efficient and interpretable models that can be more easily integrated into traffic regulation systems.

2.4 Hybrid and Optimized Models

Hybrid and optimization-based models have appeared as effective approaches to advance traffic congestion prediction by tackling the limitations of standalone machine learning and deep learning techniques. These models combine different learning methods or integrate optimization methods to improve performance and stability.

Often, hybrid methods use neural networks together with metaheuristic optimization techniques e.g. genetic algorithms or swarm-based methods. This combination helps tune parameters better and train models more efficiently. It also makes the models more reliable, especially when working with noisy or unpredictable traffic data.

Hybrid models that use different deep learning components, such as CNN-RNN or CNN-LSTM, have shown good results in traffic forecasting. They have also been effective for forecasting and managing traffic signals (Islam et al., 2022). By combining spatial feature extraction with temporal sequence modeling, these models can capture complicated traffic patterns more effectively than single-structure models.

Overall, hybrid and optimized models provide a flexible and effective framework for congestion prediction. This makes them appropriate for applications that require high accuracy and trustworthiness, such as responsive traffic signal control systems (Cheng et al., 2022; Topilin et al., 2025).

2.5 Use of Traffic Datasets

How well traffic congestion forecasting models perform depends mostly on the quality and relevance of the data used for training and evaluation (Medina-Salgado et al., 2022). Most research uses real-world traffic data from working transportation systems, such as loop detectors, smart traffic signals, and IoT-based sensors. These large-scale datasets are frequently analyzed using big data techniques to improve traffic prediction performance (Zhu et al., 2019).

Common data sources include large-scale traffic databases and open-data platforms provided by transportation authorities. Publicly available datasets are also widely used in research, as they offer accessible and standardized data for model development and comparison.

In this study, a real-world traffic dataset obtained from the Kaggle platform is used for congestion prediction and analysis. Specifically, the Traffic Prediction Dataset provides structured traffic-related data that can be used to model and forecast traffic conditions. These datasets include important temporal information that supports the development of machine learning and deep learning models.

Data preprocessing is a key step in preparing datasets for analysis. This process regularly means managing missing values, correcting inconsistent records, removing outliers, and including useful time-based features such as time of day or day of week. These steps make the data better reflect real traffic patterns and support machine learning and deep learning.

2.6 Research Gaps and Opportunities

AI-based traffic congestion forecasting has made a lot of progress, but there are still some limitations in the research. Many models mostly use historical data and do not make full use of real-time traffic information, which makes them less effective for real-world traffic management (Liu et al., 2025).

In addition, many studies focus on isolated intersections or short road segments. Because of this limited scope, models struggle to work well across connected urban traffic networks, where congestion often spreads from one intersection to another.

Another important challenge is that advanced models can be very complex. Hybrid and graph-based deep learning methods often predict well, but using them in real-world situations can be difficult because they need a lot of computing power and are hard to interpret (Kashyap et al., 2022). These issues can make them less practical for traffic operations.

These challenges show there are valuable areas for future research. Researchers increasingly need models that are accurate, scalable, easy to understand, and ready for real-time use. Another key area is combining congestion prediction with responsive traffic signal control, which helps move traffic management from reacting to problems to preventing them.

2.7 Historical Development of Traffic Congestion Forecasting

Early studies on traffic congestion forecasting mostly used statistical and analytical methods. Researchers often use techniques such as linear regression, ARIMA, and likelihood-based models to analyze historical traffic data and identify patterns, especially during busy periods. While these approaches provided useful baseline insights, they were limited in representing nonlinear relationships and rapid temporal changes under traffic conditions.

With advances inside computational methods, machine learning techniques began to play a more important role in traffic forecasting. Models like decision trees, support vector machines (SVM), k-nearest neighbors (KNN), and logistic regression offered more flexibility because they could learn complex relationships directly from traffic data. These methods predicted outcomes better than traditional statistical approaches, but they still struggled to model long-term patterns over time.

Recent progress in artificial intelligence has improved how accurately we can project traffic congestion. Recurrent neural networks (RNNs), such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models, are good at

finding patterns in traffic data over time. In addition, spatiotemporal and graph-based neural networks have made it possible to model both spatial interactions and temporal dynamics inside traffic networks.

Overall, the evolution of traffic congestion forecasting shows a clear shift from traditional statistical methods to more advanced, data-driven approaches that can better handle the complexity of modern urban traffic systems.

2.8 Previous Studies on Traffic Congestion Forecasting

Recent studies show that artificial intelligence methods are effective for predicting road congestion and traffic flow in different situations. Katambire et al. (2023) compared ARIMA and LSTM models for short-term traffic flow prediction at an intersection and found that LSTM performed better than ARIMA in terms of MAE, MAPE, and RMSE. The results show that learning from time-based patterns is important for predicting congestion. Muhammed et al. (2024) presented a hybrid model that uses Graph Attention Network, Bi-directional LSTM, and Attention (GAT-Bi-LSTM-Attention) to estimate traffic speeds at different times with freeway datasets. The study demonstrated that incorporating spatial interactions alongside time attention components markedly increased forecasting precision as compared with simpler deep learning structures.

Tree-based ensemble algorithms have also shown strong performance in congestion-related prediction tasks. Chen et al. (2021) used an XGBoost model to predict travel time with real-world probe-vehicle data. They found it was more accurate and efficient than traditional gradient boosting models. This suggests that ensemble learning works well for complex and noisy traffic data.

More advanced hybrid and physics-informed approaches have recently gained attention. Li et al. (2025) created a physics-informed Deep Operator Network (DeepONet) that uses traffic flow preservation principles during training. They found that this approach improved forecast accuracy and better captured how congestion spreads across space and time.

Sakib (2025) also tested several tree-based machine learning models, such as Decision Trees, Random Forest, Gradient Boosting, XGBoost, and LightGBM, in very crowded urban traffic conditions. The study found that Random Forest achieved the best overall performance, while explainable AI tools such as SHAP

provided important knowledge of the factors influencing congestion, thus increasing model explainability.

Overall, these studies indicate a clear trend toward hybrid, spatiotemporal, and interpretable AI-based models for traffic congestion forecasting. Most current research looks at forecasting performance, but does not focus much on how congestion predictions can be used in real traffic control, such as adaptive signal timing. This shows a need for systems that connect accurate prediction with practical traffic management. While a wide range of models such as ARIMA, LSTM, XGBoost, as well as graph-based neural networks have been successfully applied in traffic congestion prediction, the implementation of all these models is beyond the scope of this study.

Instead, this research focuses on evaluating a representative set of machine learning models and selecting the most suitable approach based on performance, computational effectiveness, and practical usefulness.

The reviewed models are included to offer a complete comprehension of the current state of the art and to position the proposed approach within the wider research context.

2.9 Comparative Summary Table

The following table presents a comparative summary of representative quantitative findings from recent studies on traffic congestion forecasting. It stresses the diversity of modeling methods applied in the literature and summarizes their reported performance and key contributions.

Study	Model	Key Findings
Katambire et al.(2023)	LSTM vs. ARIMA	Compared ARIMA and LSTM for short-term traffic-flow forecasting at Muhima Junction (Rwanda). Using MAE, MAPE and RMSE as evaluation measures, this LSTM model attained lower errors than ARIMA and was selected as the best-fitting model for monthly traffic-flow prediction, supporting its use for adaptive traffic-signal control.
Muhammed et al. (2024)	GAT-Bi-LSTM-Attention	Proposed a hybrid GAT-Bi-LSTM-Attention model for multi-horizon traffic-speed prediction on PeMSD4/PeMSD7 freeway datasets. The full model (GAT-Bi-LSTMA) achieved MAE ≈ 0.19 , RMSE ≈ 0.43 and MAPE $\approx 9.7\%$ on PeMSD4—substantially better than ablated variants without Bi-LSTM, attention or GAT—showing the benefit of combining graph-based spatial representation with bi-directional LSTM and time-based attention.

Topilin et al. (2025)	CPO–CNN–LSTM–Attention (Hybrid Deep Model)	Developed a Crested Porcupine Optimizer (CPO)-CNN-LSTM-Attention model for urban traffic-flow forecasting on Paris road data. The combined architecture achieved RMSE in the range 17.35–19.83 and MAE about 13.98–14.04, outperforming comparison models and demonstrating strong short-term prediction performance for congestion-prone urban corridors.
Chen & Fan (2021)	XGBoost (Tree-Based Gradient Boosting)	Used an XGBoost model to predict freeway travel time from probe-vehicle data. After hyperparameter optimization, XGBoost delivered higher prediction effectiveness as well as efficiency than a standard Gradient Boosting model, indicating that boosted decision-tree ensembles are effective for congestion-related travel-time forecasting on freeways.
Li et al. (2025)	Physics-Informed DeepONet (PI-DeepONet)	Introduced a physics-informed Deep Operator Network for traffic-state estimation, embedding the traffic-flow conservation law and the Greenshields fundamental diagram into the learning process. Experiments on real-world traffic data showed that DeepONet and PI-DeepONet improved MAE and RMSE by up to about 24.17% and 27.84%, respectively, compared with state-of-the-art baselines, while better capturing congestion propagation in space and time.
Sakib (2025)	Tree-Based ML + XAI (DT, RF, GBM, XGBoost, LightGBM + SHAP)	Evaluated five tree-based models for travel-time prediction amid heterogeneous, highly congested traffic conditions in Dhaka (Bangladesh) using Google Maps travel-time data. Random Forest achieved the lowest RMSE, MAE and MAPE and the highest R^2 among the models. SHAP analysis showed that “travel time 15 minutes prior” was the most influential predictor, followed by earlier travel times and section length, highlighting the value of explainable ensemble models for congestion-related travel-time prediction.

Comparative Table 1

2.10 Summary

This chapter reviewed the main research on traffic congestion forecasting, showing how methods have developed from early statistical and machine learning techniques to more advanced deep learning, hybrid, and physics-informed models. The reviewed studies show that artificial intelligence–based techniques generally outdo standard methods in observing the temporal and spatial dynamics of urban traffic infrastructures.

The literature also shows a clear shift toward spatiotemporal and combined models, such as CNN–LSTM, graph-based neural networks, and optimization-based approaches, which offer increased forecast exactness and stability. In addition, recent studies show the growing importance of model explainability

through explainable AI techniques, helping to better understand the factors influencing congestion.

Although there has been progress, current research still has some limitations. These include not using enough live traffic data, not fully considering how congestion spreads across networks, and not connecting congestion prediction closely enough to real-world traffic signal control strategies. These gaps point to the need for forecasting systems that are not simply accurate, but equally scalable, interpretable, and relevant for actual applications.

Building on these conclusions, this study centers on developing a congestion prediction framework customized to real traffic signal data and consistent with the requirements of adaptive signal control. The next chapter builds based based based on this foundation by describing the suggested methodology, dataset attributes, feature engineering process, and model development steps used to address the identified research gaps.

Chapter 3 Proposed Approach

3.1 Overview

This chapter explains the method followed to develop a traffic congestion prediction system and boost traffic signal efficiency. The process starts with understanding the dataset and preparing it for analysis, followed by feature extraction, model training, and evaluation. Finally, an responsive traffic signal mechanism is introduced based on the model projections.

3.2 Dataset Description

The dataset used in this project was obtained from Kaggle (Fedesoriano, 2021) and contains traffic flow data collected from multiple road junctions over time. It includes 48,120 records representing hourly vehicle counts.

Each observation relates to a specific junction at a specific time, permitting the model to learn temporal traffic patterns.

The dataset contains the following main attributes:

DateTime: Represents the timestamp of each observation

Junction: Indicates the junction number (1 to 4)

Vehicles: Number of vehicles recorded at that time

The dataset contains **4 classes** corresponding to the four junctions.

The dataset is suitable for time-based prediction tasks, as it reflects real traffic behavior such as peak hours and daily variations.

Dataset Sample

DateTime	Junction	Vehicles	ID
2015-11-01 00:00:00	1	15	20151101001
2015-11-01 01:00:00	1	13	20151101011
2015-11-01 02:00:00	1	10	20151101021
2015-11-01 03:00:00	1	7	20151101031
2015-11-01 04:00:00	1	9	20151101041
2015-11-01 05:00:00	1	6	20151101051
2015-11-01 06:00:00	1	9	20151101061
2015-11-01 07:00:00	1	8	20151101071
2015-11-01 08:00:00	1	11	20151101081
2015-11-01 09:00:00	1	12	20151101091
2015-11-01 10:00:00	1	15	20151101101
2015-11-01 11:00:00	1	17	20151101111
2015-11-01 12:00:00	1	16	20151101121
2015-11-01 13:00:00	1	15	20151101131
2015-11-01 14:00:00	1	16	20151101141

Figures 1

3.3 Data Preprocessing

Before training the model, the dataset was prepared to guarantee consistency and usability.

First, the DateTime column was converted into a proper datetime format. After that, the dataset was sorted chronologically to maintain the time sequence.

New time-related features were then extracted from the DateTime column, including:

Hour of the day

Day of the week

These features help the model understand traffic movements more effectively.

Additionally, the dataset was checked for missing values and inconsistencies. Since the dataset was relatively clean, only minimal preprocessing was required.

3.4 Feature Engineering

To improve the forecasting capability, additional features were created based on the existing data.

Time-based features such as hour and day of week were used as key inputs to capture traffic patterns.

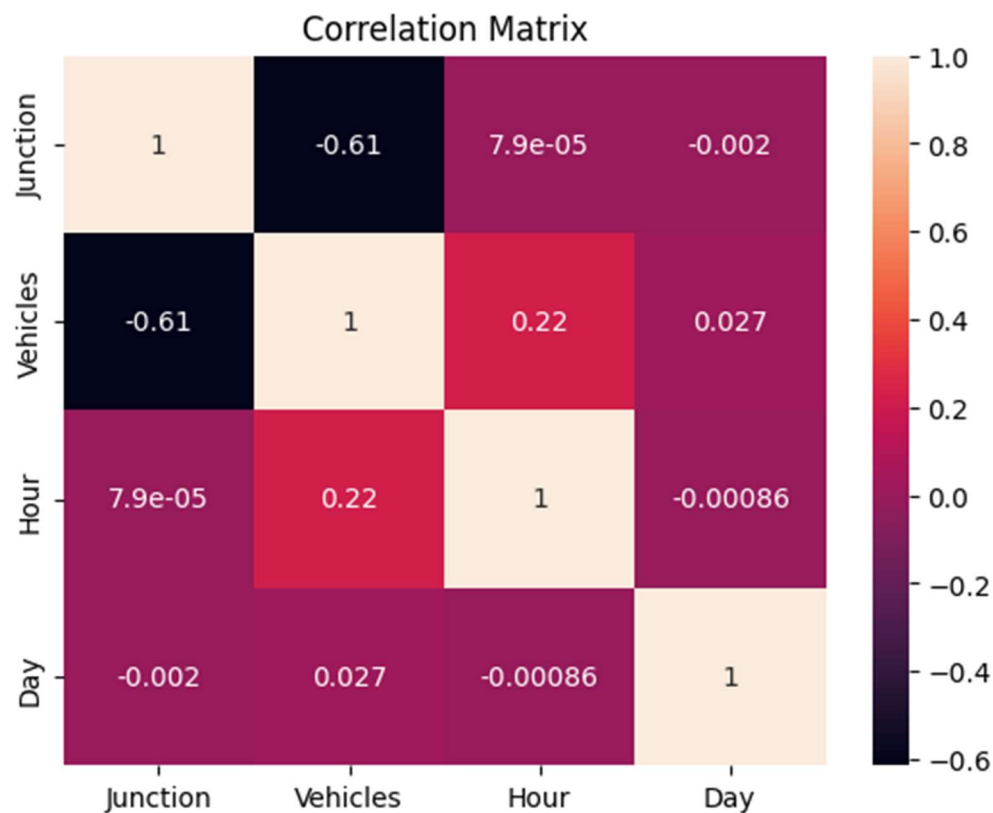
A new target variable called congestion was also created based on the number of vehicles. A threshold was used to classify traffic conditions into:

0 → No congestion

1 → Congestion

This transformation allowed the problem to be treated as a classification task. Although the dataset contains multiple junctions, the Junction variable was included as an input feature to allow the model to learn location-specific traffic patterns. However, this approach could make it more difficult for the model to generalize. The model may begin to link traffic jam patterns to specific junctions instead of showing general traffic behavior. To address this, researchers could create separate models for each junction, permitting them to focus on local

traffic trends. This method is suggested for future studies and could lead aimed at more accurate forecasts.



Figures 2

3.5 Model Selection

This study looked at four machine learning models to predict short-term traffic congestion: logistic regression, random forest, support vector machine (SVM), and

gradient boosting. The models were selected to show different methods and to compare their prediction accuracy and efficiency.

The Logistic Regression model was implemented with `max_iter = 2000`, `class_weight = "balanced"`, and `random_state = 42`.

The Random Forest model was implemented with `n_estimators = 300`, `max_depth = 10`, `class_weight = "balanced"`, and `random_state = 42`.

The Gradient Boosting Classifier was implemented with `n_estimators = 200`, `learning_rate = 0.05`, `max_depth = 3`, and `random_state = 42`.

The SVM model was implemented using the RBF kernel with `probability = True`, `class_weight = "balanced"`, and `random_state = 42`.

Standardization was applied for Logistic Regression as well as SVM through a preprocessing pipeline.

Gradient Boosting was chosen as the final model because it offered a good balance of predictive accuracy, computational efficiency, and fit for structured traffic data. Compared to the other models tested, it delivered strong classification results and kept training and prediction times reasonable. In addition, tree-based ensemble algorithms are well suited for tabular datasets with temporal and lag-based features, as they can represent nonlinear relationships without demanding extensive feature transformation.

I tested the stability of the models using time-series cross-validation with five splits. This approach let us see whether the models performed consistently over different periods in the dataset. The results showed that Gradient Boosting was both reliable and effective for this study.

In this study, we did not use extensive optimization methods such as grid search or random search to set model parameters. Instead, we selected hyperparameter values based on practical testing and by using common settings for tree-based ensemble models. The main objective of the study was to evaluate the feasibility of machine learning for congestion prediction and responsive traffic signal support rather than to obtain the absolute maximum possible accuracy via comprehensive parameter refinement.

The Gradient Boosting model showed good accuracy, precision, recall, and trustworthiness during temporal validation. Future research could use systematic

tuning methods to improve predictions and compare those results with this approach.

index	Model	Accuracy	Precision	Recall	F1 Score	ROC.AUC
0	Gradient Boosting	0.9802515331046668	0.9662387066096053	0.9446768944676894	0.9553361542078044	0.9966390900655153
1	Random Forest	0.976093961126702	0.9332431213351375	0.9618781961878197	0.9473443223443223	0.9966813480483223
2	SVM	0.9689221494647126	0.8933729821580288	0.9776847977684798	0.9336293007769145	0.9943392351367348
3	Logistic Regression	0.9461594426853549	0.8148862321635172	0.9823337982333799	0.8908094435075885	0.9946618645665879

Figures 3

Additionally, the execution time of the models was compared to evaluate their processing effectiveness.

*** مقارنة زمن التنفيذ بين النماذج:

	Model	Training Time (sec)	Prediction Time (sec)
0	Logistic Regression	0.478001	0.006330
1	Random Forest	6.654170	0.152570
2	Gradient Boosting	9.290283	0.016335
3	SVM	35.590247	1.392408

Figures 4

3.6 Training Process

I divided the dataset into training and testing sets. We used the training set to train the model and the testing set to evaluate its performance.

During training, the model learned the relationship between the input variables (such as hour, day, and junction) and the target variable (congestion).

The training process is intended to produce a model capable of making dependable predictions on unseen data.

3.7 Model Evaluation

After training, the model was evaluated using the testing dataset.

The evaluation focused on common classification standards such as:

Accuracy: 0.968

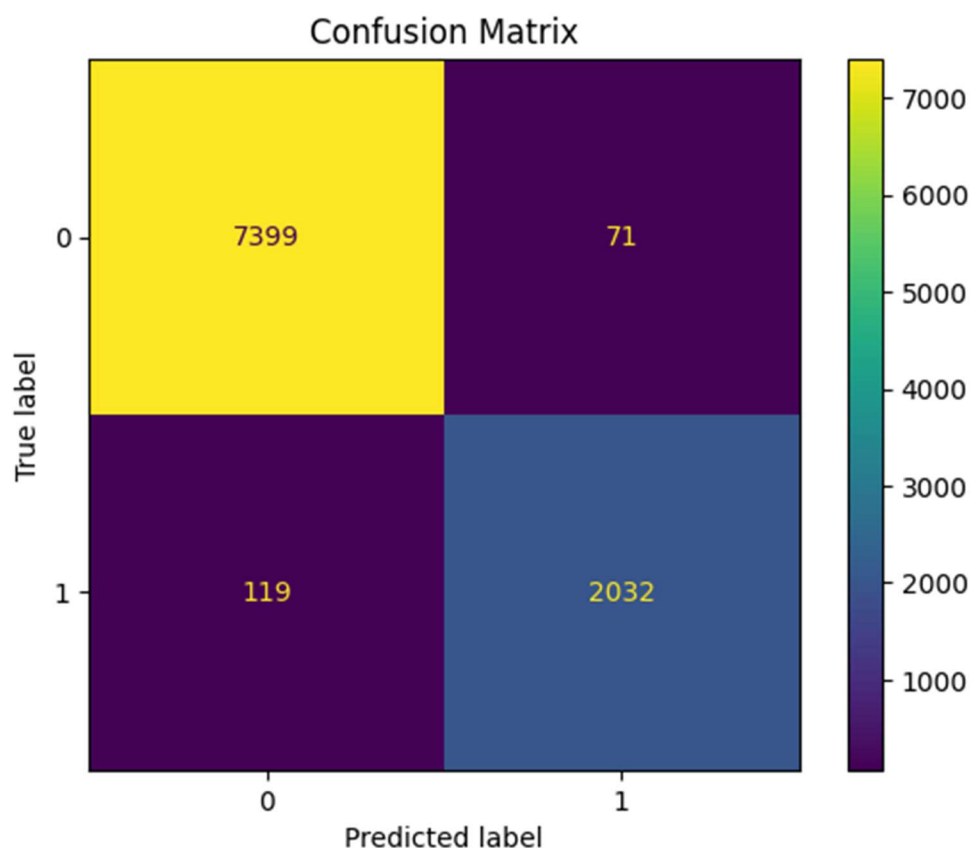
Precision: 0.893

Recall: 0.977

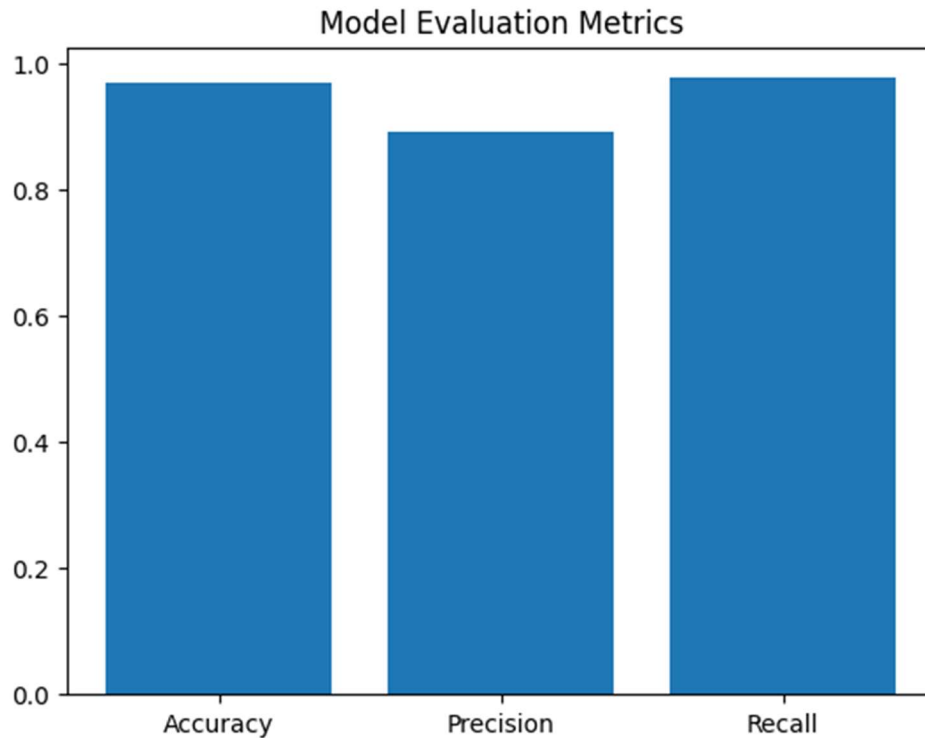
F1-score: 0.933

These metrics were used to measure how well the model predicts congestion.

The results showed that the model can reasonably capture traffic patterns as well as predict congestion levels, rendering it suitable for further application. In addition to classification metrics, training time and prediction time were compared across the evaluated models for assessing the models' computational practicality for traffic management applications.



Figures 5



Figures 6

3.8 Adaptive Traffic Signal Mechanism

After obtaining the prediction results, an responsive traffic signal mechanism was implemented to emulate real-life application.

If the model predicts congestion in the next hour, the system adjusts the traffic signal timing by increasing the green light duration. The adjustment depends on the prediction probability, allowing the system to respond more effectively to expected traffic conditions.

If no congestion is predicted, the signal remains in its normal timing mode.

This mechanism demonstrates how anticipatory models can be integrated with traffic regulation systems to reduce congestion and boost traffic flow.

3.9 Summary

This chapter presented the proposed approach used in this project, starting from data preparation and feature engineering to model training and evaluation. It also

introduced an responsive traffic signal mechanism based on prediction results. The next chapter will present the experimental results and discussion.

Chapter 4: Experimental Results and Discussion

4.1 Overview

This chapter presents the results obtained from the implemented model and discusses its performance in predicting traffic congestion. It also evaluates the effectiveness of the responsive traffic signal mechanism based on the model projections.

4.2 Model Performance

Gradient Boosting Classifier

The Gradient Boosting Classifier was trained using the prepared dataset and evaluated on the testing data. The model showed good ability in the capturing of traffic patterns and predicting congestion.

The evaluation was based on common classification metrics, including:

Accuracy: Measures the overall correctness of predictions

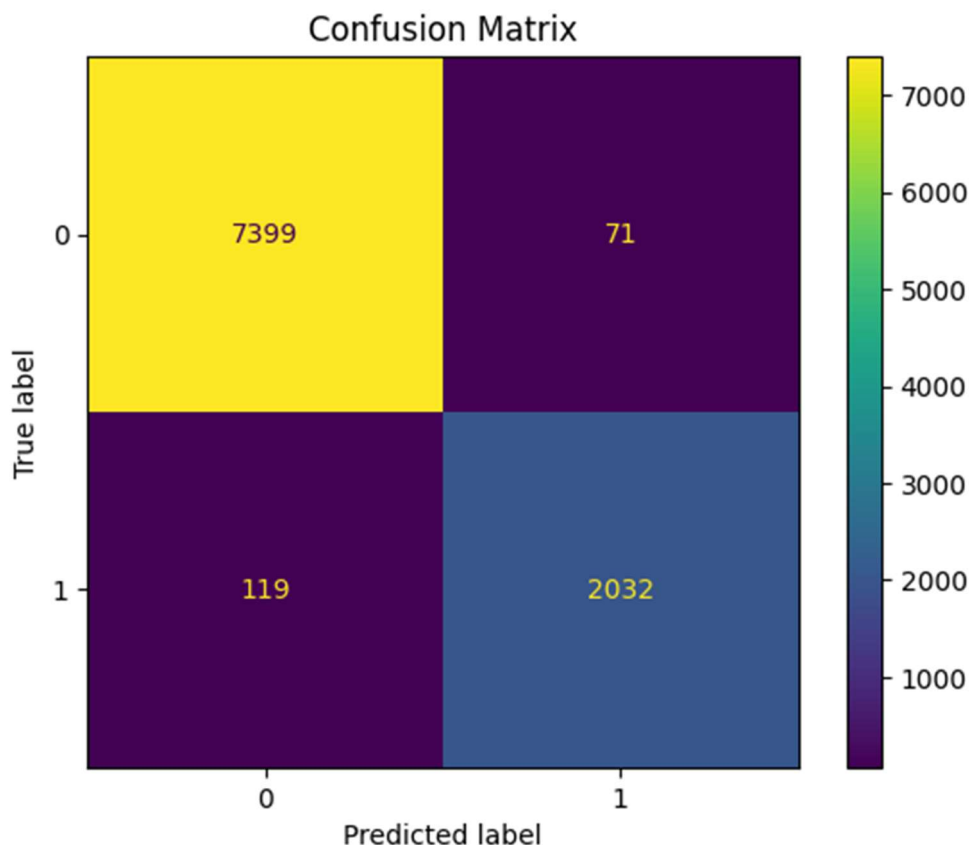
Precision: Measures how many predicted congestion cases were correct.

Recall: Measures how many actual congestion cases were correctly identified.

F1-score: Provides a balance among precision & recall, and is especially useful for imbalanced classification problems. The model attained a high F1-score, denoting a good balance between detecting congestion cases and avoiding false alarms. This suggests that the model performs reliably in identifying congestion scenarios free of excessive misclassification. The results show that the model can clearly distinguish between congested and non-congested traffic conditions. This suggests that the selected features, particularly time-based features such as hour and day of the week, play an important role in predicting traffic behavior. To better assess how well the model works, we looked at its confusion matrix. The results showed many true positives and true negatives, which means the model can correctly identify both congested and non-congested traffic conditions.

The number of false negatives (missed congestion cases) was relatively low, which is particularly important in traffic management applications, as failing to detect congestion may cause inefficient signal control decisions.

Similarly, the number of false positives was limited, suggesting that the model does not frequently trigger unnecessary congestion alerts.



Figures 7

4.3 Analysis of Predictions

The results suggest that traffic congestion is more common at certain times, especially during peak hours.

For instance, we saw higher congestion levels at these times:

Morning hours (7 AM to 9 AM)

Evening hours (4 PM to 7 PM)

This pattern is similar to what we see in real life, when people are traveling to and from work.

Additionally, the model showed consistent predictions throughout different junctions, although some variations were observed depending on traffic intensity at each location.

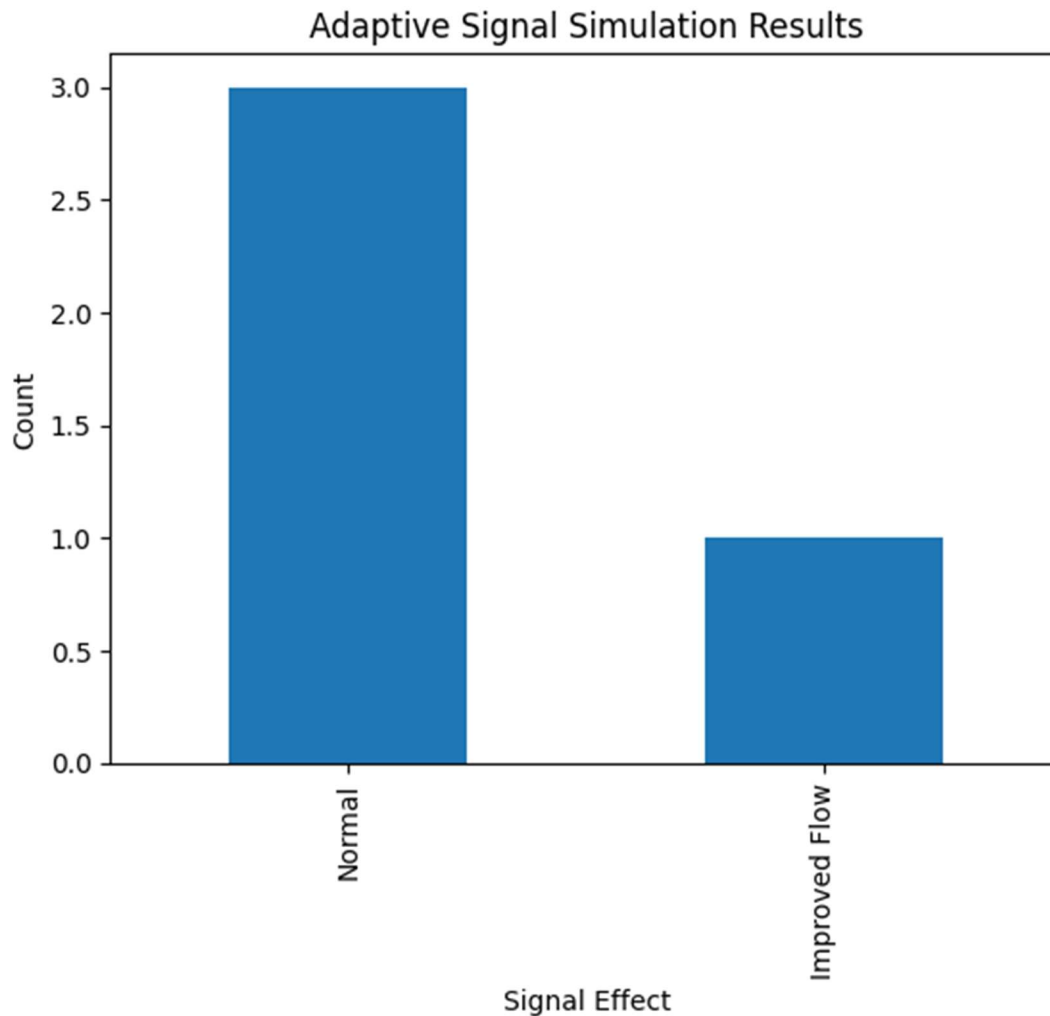
4.4 Adaptive Traffic Signal Evaluation

To evaluate the practical impact of the system, an responsive traffic signal mechanism was applied based on the predicted congestion.

When congestion was predicted, the system increased the green signal duration. This tweak helped simulate augmented traffic movement by allowing more vehicles to pass through the intersection.

The adaptive mechanism showed the following advantages:

- Reduction in predicted congestion impact
- Improved signal responsiveness
- Better utilization of green time



Figures 8

Figure X: Intelligent Traffic Signal Simulation Results

The simulation results show that a significant portion of predicted congestion cases were associated with extended green signal timing, implying a possible improvement in traffic flow under the adaptive signal control strategy.

- Reduced potential vehicle accumulation at intersections
- Smoothed traffic flow flow during peak periods
- More efficient use of signal timing compared to fixed timing systems

Although this evaluation is based on simulation, it demonstrates how predictive systems can be integrated with traffic control systems to support smarter traffic management.

4.5 Discussion

To evaluate how well the proposed model works, its results can be compared to a basic baseline that uses historical averages. This baseline predicts congestion by looking at regular patterns at certain times, such as peak hours, without using machine learning.

The Gradient Boosting model gives more accurate and flexible predictions than this baseline because it can find complex patterns and changes over time in traffic data. This helps the model spot congestion trends that simple historical rules might miss.

As a result, the proposed approach clearly improves on standard methods when predicting short-term traffic congestion. Previous research shows that LSTM models are good at capturing time patterns in traffic data, and ARIMA models are often used as statistical baselines.

Compared to these models, Gradient Boosting provides a good balance of forecast precision, efficiency, and ease of use. Tree-based models like Gradient Boosting also work well with structured datasets that have engineered features, which fits the data used in this study. This makes the approach practical for actual traffic infrastructures.

The results show that machine learning can be used effectively to forecast traffic congestion.

One main strength of this method is its simplicity. By using a structured dataset and time-based features, the model performed well without needing complex inputs like weather or sensor data.

There are some limitations to consider:

The dataset contains only a small number of junctions.

External components such as weather and accidents were not included.

In spite of these limitations, this approach delivers a strong starting point for future improvements and practical uses.

4.6 Summary

This chapter presented the experimental results and analysis of the proposed system. The Gradient Boosting model demonstrated good performance in predicting traffic congestion, and the responsive traffic signal mechanism showed encouraging potential in boosting traffic flow.

Chapter 5: Conclusion and Future Work

5.1 Introduction

This study proposed a machine learning-based system for predicting traffic congestion at signalized intersections and supporting responsive traffic signal control.

The main research question investigated whether a Gradient Boosting model could outperform a simple historical average baseline in predicting short-term traffic congestion. The experimental results show that the Gradient Boosting model performed well, with high accuracy, precision, recall, and F1-score. This means it can reliably detect traffic congestion patterns. The model also outperformed a baseline based on historical averages by capturing nonlinear relationships and changes over time in the traffic data. This shows that machine learning techniques can provide more accurate and flexible congestion predictions than traditional rule-based methods. In addition, an responsive traffic signal mechanism was implemented and evaluated through a simulation-based approach. The results indicated that increasing green signal duration during predicted congestion periods can help reduce vehicle accumulation and boost traffic traffic traffic operation efficiency. Although the adaptive control was evaluated in a simplified simulation environment, the findings show the feasible effectiveness of integrating prediction models with traffic signal control strategies.

Overall, the study shows that Gradient Boosting can effectively forecast road congestion and that prediction-driven signal control can contribute to boosting traffic management performance. The findings of this study directly answer the research question. The results show that the Gradient Boosting model outperforms a simple historical average baseline in predicting short-term traffic congestion. This demonstrates the effectiveness of machine learning approaches in capturing complex traffic patterns and supports their use in intelligent traffic management systems.

5.2 Future Work

Although the proposed approach attained promising results, there are several opportunities for improvement and further development.

First, additional data sources can be incorporated to improve forecast precision. For example, including weather conditions, road events, or special events may yield a more complete understanding of traffic behavior.

Second, more advanced models such as deep learning techniques can be explored. Models like Long Short-Term Memory (LSTM) are particularly suitable for time-series data and may improve long-term traffic prediction.

Third, the responsive traffic signal system can be extended to operate in real-time environments. Integrating the model with live traffic data and actual signal controllers would make the system more practical and influential.

Finally, the system can be expanded to cover a larger number of junctions and different cities, which would improve its scalability and wider applicability.

References

1. FEDESORIANO , Traffic Prediction Dataset
<https://www.kaggle.com/datasets/fedesoriano/traffic-prediction-dataset>
2. Medina-Salgado, et al. (2022). Urban traffic flow prediction techniques: A review. *Sustainable Computing: Informatics and Systems*, 35, 100739.
<https://doi.org/10.1016/j.suscom.2022.100739>
3. Rahmani, et al. (2023). Graph neural networks for intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 24(8), 8846–8859. <https://doi.org/10.1109/TITS.2023.3257759>
4. Katambire, et al. (2023). Forecasting the traffic flow by using ARIMA and LSTM models: Case of Muhima Junction. *Forecasting*, 5(4), 616–628.
<https://doi.org/10.3390/forecast5040034>
5. Chen, et al. (2021). A freeway travel time prediction method based on an XGBoost model. *Sustainability*, 13(15), 8577.
<https://doi.org/10.3390/su13158577>
6. Cheng, et al. (2022). Combination predicting model of traffic congestion index in weekdays based on LightGBM–GRU. *Scientific Reports*, 12, 2912.
<https://doi.org/10.1038/s41598-022-06975-1>
7. Sakib, N. (2025). Machine learning-based travel time prediction models for complex traffic conditions. In *Proceedings of the 7th International Conference on Advances in Civil Engineering (ICACE 2024)*. SSRN.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5238355
8. Zhu, et al. (2024). *Adaptive graph attention and long short-term memory network for traffic flow prediction*. *Mathematics*, 12(2), 255.
<https://doi.org/10.3390/math12020255>
9. Topilin, et al. (2025). *Traffic flow prediction via a hybrid CPO-CNN-LSTM-Attention architecture*. *Smart Cities*, 8(5), 148.
<https://doi.org/10.3390/smartcities8050148>
10. Wang, et al. (2021). *Learning the solution operator of parametric partial differential equations with physics-informed DeepONets*. *Science Advances*, 7(40), eabi8605. <https://doi.org/10.1126/sciadv.abi8605>
11. Li, et al. (2025). *Physics-informed deep operator network for traffic state estimation*. arXiv preprint. <https://arxiv.org/abs/2508.12593>

12. Liu, et al. (2025). *A review of traffic flow prediction methods in intelligent transportation system construction. Applied Sciences*, 15(7), 3866.
<https://doi.org/10.3390/app15073866>
13. Kashyap, et al. (2022). Traffic flow prediction models – A review of deep learning techniques. *Cogent Engineering*, 9(1), 2010510.
<https://doi.org/10.1080/23311916.2021.2010510>
14. Zhu, et al. (2019). Big data analytics in intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 20(1), 383–398. <https://doi.org/10.1109/TITS.2018.2815678>
15. Fang, et al. (2022). Attention meets long short-term memory: A deep learning network for traffic flow forecasting. *Physica A: Statistical Mechanics and Its Applications*, 587, 126485.
<https://doi.org/10.1016/j.physa.2021.126485>
16. Wang, et al. (2021). A hybrid deep learning model with 1DCNN-LSTM-attention networks for short-term traffic flow prediction. *Physica A: Statistical Mechanics and Its Applications*, 583, 126293.
<https://doi.org/10.1016/j.physa.2021.126293>
17. Islam, et al. (2022). Using CNN-LSTM to predict signal phasing and timing aided by high-resolution detector data. *Transportation Research Part C: Emerging Technologies*, 141, 103742.
<https://doi.org/10.1016/j.trc.2022.103742>
18. Lv, et al. (2015). Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2), 865–873.
<https://doi.org/10.1109/TITS.2014.2345663>
19. Vlahogianni, et al. (2014). Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies*, 43, 3–19.
<https://doi.org/10.1016/j.trc.2014.01.005>
20. Genders, et al. (2019). Using a deep reinforcement learning agent for traffic signal control. arXiv preprint.
<https://arxiv.org/abs/1904.08117>